Investigating and Modelling Rationale Style Arguments

Stubbings, Georgina

Award date:
2015

Awarding institution:
University of Bath

Investigating and Modelling Rationale Style Arguments

Georgina Faye Stubbings

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Computer Science

April 2015

Table of Contents

Acknowledgements

Any major project has influences from many areas of life, and especially the people in it to help get the job done. However, there are certainly some significant contributors and I would like to personally thank them here.

Firstly, I would like to thank my family, especially my mother, Gillian Stubbings, whose unwavering faith that something like this could be achieved has been an invaluable resource.

The thesis also owes its existence in part to my father, Christopher George Stubbings, who (from wherever he is now) still holds a driving influence on my desire to try, and keep trying in the face of the many ups and downs that this process has taken.

I would like to thank my fiancé Ben Evans, whose incredible support has given me the much needed space and contemplation with which to take this on. Not to mention the steady provision of chocolate, escapism and lemonade.

Finally, I would like to thank my supervisory team and particularly Peter Johnson, whose continued belief in my journey, careful guidance and impressive ability to draw the best out of his students has led me here. This process has encouraged me to grow in ability and confidence. I cannot thank him enough.

I would also like to thank the following people who have helped me through the process in one way or another; Tim and Sarah Evans (for everything), Fiona Anderson (for the Sambuca and joy), Charmaine Capener (for the dancing and venting), Ade Anderson (for the hospitality and food related distractions), Dave Evans (for the gambling and random carbohydrates), Kate and Jessica Mansfield (for being hilarious), Ian Fairholm (for the joy of teaching), Claire Snaydon (for facilitating contact with the outside world), Villiers Park Educational trust (for the willing students and unforgettable experiences, and finally, all of my patient and (now) widely distributed friends.

Abstract

This thesis investigates how the intended direction (either self or other) and perception of future use when constructing a rationale style argument can impact upon decision confidence and argument quality (in terms of rhetorical structure and the use of rebuttals). The literature review reveals emerging needs for further understanding of how the perception held of intended rationale direction can impact on the attitudes held about the decision and structure of the rationale. Rationale style arguments were the focus of investigation due to their prevalence in research, potentially rich and varied argumentative structures and wide scope of utility in other domains. The findings inform a rationale style argument model that assists in scoping the argument context, adding further dimensions including the intended direction of the rationale (self or other) and argument competency. The thesis proposes two new frameworks that offer a semi-automated solution to argument quality analysis. A good level of agreement between the new quality analysis frameworks and the original Toulmin based quality scheme used was found and the utility of the findings for future feedback tools and online argument analysis is discussed. The new semi-automated frameworks would enable analyses to be carried out rapidly and with less subjective judgement. The work may also have applications for educational tool designs that seek to incorporate argument analysis and feedback on text based arguments.

Part One: Introduction and Overview

1.1 Introduction

This thesis is concerned with the investigation of rationale style arguments. These types of rationale contain argumentative components and explanations to support the claims made within. Argument is considered from a purpose perspective, with statements categorised in terms of whether they offer support for a claim or refute an opposing claim. This broad categorisation is used in combination with an analysis of some of the structural discourse features, such as the use of 'but' and 'however' to indicate an alternative view.

The empirical work examines how the features of these types of arguments, such as the use of supporting information and considerations of alternatives, vary depending on whether the author holds a perception of the rationale as being written as self or other directed. In addition, the empirical work will examine if the intended direction of a rationale may impact decision making in terms of the confidence held in the decision. A beneficial impact of constructing a rationale as part of a task, that of enhanced information recall, is also empirically studied to determine whether this too is influenced by the perception of direction. The thesis utilises various methods for argument analysis and quality evaluation and proposes improved quality analysis frameworks based upon the findings.

1.2 Research Questions

This thesis empirically investigates the following research questions:

1. Does the perception of direction (either self or other directed) and future use held by an author when constructing a rationale style argument influence perceived decision confidence?

2. Does the perception of direction (either self or other directed) and future use held by an author when constructing a rationale style argument influence the structures within and the quality of arguments?

3. Does the perception of direction and future use held by an author when constructing a rationale style argument influence engagement with task material and thus recall of new information?

4. Does the length and structure of the rationale style argument have any bearing on perceived confidence in a decision based on the rationale?

5. Can the perception of direction held by an author of a rationale be manipulated using a written prompt during a decision making task?

6. Do expert and novice authors differ in their attitude and approaches to rationale construction in terms of the use of knowledge telling and knowledge transforming strategies and are these reflected in the quality and use of measurable argumentative strategies?

7. Can rationale style arguments be modelled in terms of expected linguistic and argument structure?

8. Can Rhetorical Structure Theory (and automated text analysis procedures) be empirically mapped onto the Toulmin model of argument?

9. Can Rhetorical Structure Theory and automated argument analysis tools be adapted to inform frameworks that assist in the evaluation of argument quality (in terms of the use of rebuttals)?

1.3 Thesis Overview

The thesis is broadly divided into six parts. This first part of the thesis presents the research questions, followed by the introduction and overview and finally, a statement of the thesis contributions. The second part of the thesis will discuss the theoretical background of argumentation and explanation and how these activities have been studied and demonstrated to be an important aspect of tasks and decision making. The application of the current understanding of the field of argument structure and quality analysis to the development of technological support will be examined. Following this, emerging needs drawn from the discussion will be highlighted. Part three details the body of experimental work conducted to address the research questions, through chapters 5 to 8. Part four of the thesis, chapters 9 and 10, utilise the findings in the experimental work to inform a rationale style argument model and a set of adapted argument quality analysis frameworks. Part five will draw together conclusions from across the thesis and discuss the implications of the work. This part will suggest applications and limitations along with future routes of study. Finally, part six contains the thesis Appendices. The full chapter overview is detailed below.

Chapter 1 presents the thesis introduction. This chapter encompasses the research questions, thesis contributions and a full overview.

Chapter 2 will provide a definition of rationale style argument and the relevant research into how this type of argument may have a beneficial impact on learning performance and decision confidence. These sections will examine the processes that may differ when explaining in a self or other directed context. The emerging issue of perceived rationale direction will be discussed as a result of examining the contrasts between self and other directed arguments and the potential methodological issues in the research.

Chapter 3 will describe the approaches used in research to analyse explanations and arguments firstly, in terms of structural features and secondly, the methods that identify features that may pertain to 'quality' and to determine whether these features may be related to the observed learning and confidence effects.

Chapter 4 will consider methods of argument analysis from a technology support perspective. The current systems available will be evaluated in terms of the levels of support and automated feedback and analysis they implement. The apparent need for an

argument analysis method that is intuitive, rich and accessible is discussed. The concluding statements will summarise the thesis direction in light of the full literature review.

Chapter 5 will cover an initial exploratory study using a decision scenario to prompt rationales. The trends revealed in relation to the perceived direction of the argument are outlined and an initial model of the potential impact of perceived direction on argument quality and decision confidence is proposed.

Chapter 6 adapts the methodology of the first study and adopts a more intuitive and open decision task to elicit richer rationales. The findings informed an enhanced model of how perceived direction can influence argumentative structures and how a positive attitude towards a decision may be facilitated.

Chapter 7 introduces another dimension to consider, that of expertise in argument construction. A sample of Expert arguers is contrasted with the Novice rationales gathered from the previous studies. The findings provide additional considerations for the rationale model and highlight an intriguing difference between the strategies used by Expert and Novice arguers who hold an other directed perception when constructing their rationales.

Chapter 8 demonstrates a small step in examining how rationales that vary in rhetorical structure impact upon those who receive them. A small study using a sample of the rationales is outlined. The findings help to give a perspective on how the structural aspects of an argument may relate to behaviour change and persuasion, from a receiver perspective.

Chapter 9 examines the combined findings using the analysis frameworks adopted throughout the thesis. Based on the findings a rationale style argument model is proposed that is grounded in a Toulmin style layout. The constraints that the perceived direction and expertise place on the expected argument structures are incorporated into the model.

Chapter 10 will examine the correlational relationships between the analysis approaches and how these map onto the Toulmin model of argument. Additionally, this chapter will discuss the rhetorical features of the arguments that correlate with the Toulmin quality scheme level assigned in order to inform new quality frameworks based upon a structural analysis.

Chapter 11 outlines the new adapted quality frameworks based on the findings in chapter 10. The chapter will describe the rationale behind the new quality frameworks and an analysis of agreement using the new frameworks with the original quality assessments.

Chapter 12 summarises the findings from throughout the thesis. The model of rationale style argument and the new quality frameworks are considered as having potential for applications in wider research and opportunities for further investigation are discussed. The constraints on the perceived direction effect and implications for further understanding of the previous work are discussed.

1.4 Thesis Contributions

1. This thesis proposes a theory of the 'perceived direction effect' by demonstrating in the empirical work that the perception of direction (if adequately cued) can be significantly altered using a written prompt and that this impacts upon the structure of the rationales and perceived decision confidence. The shift in perception, to that of less self directed and with a perception of future use, appears to result in the increased use of argument in the externalised rationale and an increase in perceived confidence in a decision. This conclusion is based on the findings from the first and second investigations (section 5.3.2.4, 6.3.2 and 6.3.8). Additionally, the strategies adopted by those who perceive themselves to be writing in a less self directed manner were shown to be similar to strategies used in expert arguments (see section 7.4.6). These findings have implications for research concerning the self explanation effect and rationale construction in particular.

2. The empirical work also outlines a novel rationale elicitation task (see section 6.2.3.2 for task procedure) that enables rich argument based rationales to be cued in an unstructured domain of psychological debate. The novel rationale elicitation task methodology has been demonstrated as useful for measuring new task based information recall, as an aspect of learning and to indicate task engagement. Additionally, the directional prompting procedure appears to be successful in influencing the perception of direction held by an author (see section 6.2.3.3).

3. The thesis proposes a model of typical rationale style argument (see section 9.3) in an individual context which scopes how the perceived direction effect and the additional consideration of argument expertise can impact on decision confidence and argument structure (in terms of the use of rebuttals). The model components outline expected rationale structure with regard to argumentative elements (Toulmin) and linguistic structure (rhetorical relations). This model is intended to be informative for research into rationale elicitation and support in terms of suggesting which aspects of the rationale structure need to be supported in order to enhance confidence or the depth of processing of task material. An additional contribution highlighted by the model is that the use of Contrasts in an argument may be an indicator of the presence of a rebuttal. Thus, there is a suggestion that this relation needs to be re-categorised as a presentational relation (see section 10.4 for a discussion) in the original Rhetorical Structure Theory framework. This is in order for its importance within an argument not to be overlooked, of which there is a risk of doing so with its current categorisation as a neutral, non-argumentative relation.

4. The novel use of a combination of structural text analysis procedures (Rhetorical Structure Theory, and two automated text parsers) demonstrates the utility of these approaches in the analysis of unstructured rationale style arguments. The automated parsers in particular have not been used in rationale analysis research and the use of these to detect differences in reasoning styles and argument structure appears to have been successful (see section 6.3.8.3 for an example of the PDTB text parser findings) as the findings concur with the human analysis approaches. See section 10.2 for an examination of the correlational relationships between the human and automated analysis approaches.

5. The thesis also provides a comparison of a Toulmin based argument quality analysis framework with three structural analysis approaches: Classical RST, and two automated text parsers (The PDTB and HILDA parsers). The assessment of argument quality in this respect is based on the premise that arguments with rebuttals are of a better quality than those without as they prevent circular argumentation and indicate wider consideration of the argument scope. The empirical findings enabled a novel mapping of the constructs within the analysis techniques and the Toulmin model of argument (see section 10.2.4). This offers an additional use for the structural frameworks by enabling the constructs to be categorised according to the Toulmin elements. This categorisation gives the elements an argumentative purpose that is absent from the current structural frameworks.

6. Finally, the discovery of relationships between the argument analysis methods has enabled the proposal of new quality analysis frameworks (see sections 11.3, 11.4, 11.5 for description of the new frameworks). These frameworks offer greater utility than the current structural argument analysis methods as they impose a structural hierarchy of quality that is absent from the original approaches. Additionally, they categorise structural elements in terms of argument purpose, a feature that is also omitted from the current structural analysis approaches. The original Toulmin based quality scheme upon which the new frameworks are based focusses on the use of broad categorisation of argument elements using the Toulmin model.

   The new quality frameworks proposed provide a more complex view of argument analysis than the Toulmin approach as they incorporate finer grained linguistic features in the use of relations and hence offer more detail with which to assess argument structure and quality. In addition, two of the three new frameworks produced are semi-automated in their approach which enables argument analysis to be less subjective and time consuming

and overall require less domain knowledge or expertise to use. The frameworks allow for a determination of argument quality to be made based on the output from a text based parser. The initial testing of the comparability of the new frameworks (see section 11.6) with the original quality assessment indicated that the new frameworks are potentially useful and usable for the purpose of argument quality evaluation.

The frameworks can be used to evaluate argumentative text and would be of use to researchers, educators and those in roles that require the identification and evaluation of competent arguments. Improved quality frameworks may also have the potential to form the basis of an argument feedback system which is an aspect of argument support that is currently underdeveloped.

Part Two: Literature Review

The empirical work examines how the structural components in rationale style arguments may vary as a result of the author's perceived intended direction of the argument. In order to understand the implications of this work, a discussion of the wide ranging research into explanation and argumentation (both features of rationale style arguments) will commence in chapter 2. This will include an examination of the utility of argument and explanation in both individual and collaborative contexts, with a particular focus on how these activities enhance decision making and learning. The importance of examining argument from both an individual and collaborative perspective will also be considered in section 2.3 of chapter 2. This will determine the influence that intended rationale direction may have on the structure of such arguments and how this aspect needs to be addressed in order to understand its impact on constructing a rationale on an individual level.

Chapter 3 will further consider attempts to identify, analyse and model various types of argument in research. This is considered necessary due to the argumentative elements that a rationale may incorporate. This chapter will also serve as an overview of the state of the art for argument modelling and analysis, both in terms of structure and quality and thus possible directions for research and appropriate analysis methods will emerge from this. Chapter 4 will outline the current systems available for computer supported argumentation and evaluation and will conclude the literature review.

2 Argument and Explanation: Definition and Utility

2.1 Defining Rationale Style Arguments

The literature discussed in this section will introduce the idea that rationales can be argument based, with explanation utilised as supporting evidence for the arguments made. Section 2.1.1 briefly defines the relevant terminology with regard to the current understanding of rationale and the central components of argument and explanation. It is considered useful to examine both explanation and argumentation research in order to inform appropriate analysis tools and procedures for the investigation. Following this, section 2.2 will discuss the utility of explanation and argument with regard to learning and confidence effects. Finally, section 2.3 will examine the differences in explanation and argument structure and related task outcomes that may be influenced by whether the activity is self or other directed.

The ability to argue, support, justify and defend a decision in the form of a rationale is a skill that is prevalent and indeed necessary across many domains of human behaviour and decision making. Day to day examples include police policy logs that contain rationales for on the spot decisions (Schulenberg, 2007) medical decisions that require careful evaluation of available evidence and students engaging in debate across many academic fields (Hoffman & Elwin, 2004). Software designers are also encouraged to store design rationales as part of the development process to assist in redesign and reuse (Shum & Hammond, 1994).

Rationales can be considered equally as an argumentative and a reflective, explanatory style of thought. As a result of this, rationales are utilised in many domains of research including software design, collaborative group work and education and as such the definition of a rationale may be flexible depending on the goal of the activity in which they are based. The following section will examine definitions for the two basic components of rationales, namely, explanation and argument.

2.1.1 Components of a Rationale: Explanation and Argument

Explanation and argumentation are commonly occurring types of human reasoning present in rationales that are often studied in the context of interaction and dialogue (Bex, Budzynska, & Walton, 2012). Reasoning itself refers to the explicitly expressed conclusions

and supportive statements that humans generate in response to a query or conflicting viewpoint. Explanation is usually modelled using abductive reasoning (logic), whereas argumentation is more concerned with presumptive reasoning (claims with premises).

Bex, Budzynska and Walton (2012) defined explanation in their paper as a speech act intended to "explicate why something is the case" (p.1) and the intended purpose is not to convince, or prove, but rather, to help another understand something. In contrast, the goal of argumentation is considered to be primarily to "remove (an) opponent's doubts" (p.2) about a claim or indeed convince them of its validity, by providing proof and justification.

This distinction however, is not entirely straightforward as the objective of convincing another - known as persuasion - is dependent on the receiver of the argument, not necessarily the tangible, defined structures of the text and may also be independent of the author's purpose (intended to convince or otherwise). Explanations can indeed, also be persuasive to a receiver in the correct context and have been shown to impact upon consumer behaviour and influence purchasing decisions even if not explicitly presented in an 'argument' format (Tintarev & Masthoff, 2007).

Arguments are generally considered to be statements that make claims, followed by premises upon which the claim is based. Argument can therefore, by definition, contain explanation (supporting evidence or proof) in order to strengthen the presented reasoning. Conversely, it appears explanations can also function as arguments in the correct context. To illustrate this, an example of an explanation of a product may be:

"The Philips CD player has an auxiliary input"

This statement offers a description of the functionality of the CD player. This statement can be defined as an explanation as opposed to an argument – in terms of the literature – as it helps a receiver understand the properties of the product and does not appear to be 'intended' to persuade. However, this explanation alone could be persuasive if the receiver was choosing between this and another item and this explanation refers to a key piece of decision criteria – the presence of an auxiliary output. This effect relates to the impact of receiver involvement with a message and current position with regard to the argument (Johnson & Eagly, 1989). These are key factors in establishing whether a statement will have a persuasive impact and are discussed in more depth in section 2.3.2. If this statement were

to be constructed into an obvious argument format with an externalised persuasive intent, it may appear similar to the following:

"The Philips CD player is the best option because it has an auxiliary input"

This statement makes a claim, followed by a premise upon which the claim is based. It could be argued that this statement is intended to persuade a receiver to purchase this item on this basis. However, this argument is only persuasive if receiver has an interest in this function as the argument for the Philips CD player being the 'best' is based upon this. If the receiver has no interest in this function then their view of this CD player being the best may not be altered.

The literature suggests that argument is intended to persuade and as such an argument can, as a basic function, explicate "why" something is the case, but may also contain rebuttals of an opposing view which, it could be argued, is not a necessary aspect of explanation. The following example illustrates an argument with an additional rebuttal:

"The Philips CD player is the best option because even though it does not have a sleep function, it does have a remote control for easy operation."

In this case, the statement has pre-empted an argument against the CD player and recognises that the flaw in its claim of the item being the 'best' option is the lack of a sleep timer. The rebuttal to the pre-empted argument against the CD being the 'best' based on this flaw is offered in the form of stating a compensatory function. This example helps to illustrate how argument can contain explanation to support a claim, and it is the way in which this explanation is utilised either by the author or receiver that can determine the arguments perceived structure or persuasiveness.

Both forms of reasoning have been studied widely in education, with explanation often being utilised in well-structured domains such as physics and argumentation being applied in ill-structured areas such as the social sciences. In most domains (with the possible exception of mathematics and physics) arguments may not be deductively valid, but will involve a degree of uncertainty.

The skills required to argue or explain may differ, however both are utilised and taught as acquirable skills in terms of assisting in learning conceptual knowledge and the development

of critical thinking. In comparison to explanation, the reasoning style of 'argument' may fundamentally require more skill and critical thought during construction, such as the evaluation of knowledge. In contrast, an explanation could be considered less demanding as they are most often comprised largely of knowledge statements.

Argumentation is thus, often considered as a distinct educational goal, particularly if using a constructivist approach to learning (Osborne, Erduran, & Simon, 2004a). In this regard an individual is actively involved in linking new ideas with currently held ideas, a common application of argumentation. In Contrast, Roy & Chi (2005) argued that the activity of explanation could also be trained, implying it is also, like argument, reliant on a type of acquirable skill. The research suggested there was ample evidence for an apparent division between 'good' (those who generate lengthy explanations) and poor explainers, which could be mediated with training.

A rationale is fundamentally a self-contained piece of explanation and argument. These types of arguments may in fact act as a persuader for the author, increasing confidence in a decision and the breadth of material discussed. It may be proposed that a less useful (in terms of aiding decision making or increasing confidence) rationale would contain less argument (or solely explanation) as opposed to a more complex rationale that fully justifies a decision or evaluates alternative positions (MacLean, Bellotti, & Shum, 1993). For example, a rationale with more argumentative structures may assist in group decision making by supporting consensus, reducing circular argumentation (Kuhn, 1991), prompting more critical thought and wider discussion.

As a minimum, a rationale should outline an explanation for a decision, as a rationale without an explanation would be a simple 'claim' (Xiao, 2013a). An example of a simple claim would be:

"I chose to use text based input for my software."

In order for this claim to be considered as a rationale, it would need to express an explanation for this decision, for example:

"I chose to use text based input for my software, because it is easy to implement and interact with."

The statement above is now a basic type of explanation based rationale. In order for this rationale to better explain the background for a decision, it needs to evaluate the choice in light of alternatives and any possible criticisms. This type of rationale could be more effective in group collaboration for establishing better awareness of the process that has been undertaken in order to reach the decision. It may also be more useful for a future receiver of the rationale to be aware of the possible alternatives to the main position. This is an aspect of rationale that is often studied in software design reuse, where the design rationales are captured in order to explain the features within (MacLean, et al., 1993).These aspects of a rationale that offer consideration and rejection of alternatives could be referred to as argument; the argument based rationale would be extended in this way:

"I chose to use text based input for my software, because it is easy to implement and interact with. However, I am aware that this limits the type of users and is not as easy to use as speech input, but I think in terms of hardware costs it is the right decision."

This rationale recognises a flaw in the decision and justifies this in terms of the costly alternative. This type of rationale, that produces an argument in conjunction with explanation as a basis, will be the focus of the investigation. The extent to which these extended argument aspects are more or less prevalent depending on the perception of where the rationale is directed (at the self or others) held by the author is one of the primary research questions.

In order to understand how a rationale style argument may benefit an individual in a decision making context the next section will examine how explanation and argument have been utilised to enhance confidence and learning. The mechanisms behind these observations will be discussed. In addition, research that demonstrates what may constitute a 'good' explanation or argument in terms of facilitating these benefits and which may be indicators of 'quality' will also be outlined. This will lead to an examination of the available analysis methods used to study argument quality and finally, structure.

2.2 The Utility of Explanation

2.2.1 Introduction

This section will discuss the observed benefits of eliciting explanations, including improved task performance in terms of learning and confidence as part of educational and decision making activities. In order to examine the benefits of constructing a rationale style argument

for a decision on an individual (author) level, the research into explaining in a self directed context (known as 'self-explanation') may offer insight as rationales, in terms of the previously discussed definition, may contain explanation as support for a decision.

Self explanation is a type of explanation that is commonly elicited as part of educational research. It is a reflective style of reasoning and can broadly encompass other activities such as justification and elaborated statements. Self explanation is defined (specifically in the context of the "self-explanation effect") as "the activity of explaining to oneself in an attempt to make sense of new information, either presented in a text or some other medium" (Chi, 2000, p. 163). Self explanation is considered to be a reflective activity (Chi et al, 1989), whereby the explanation generated appears to serve the purpose of clarifying the reasons behind a particular answer for the benefit of oneself. Examples of the types of explanations generated in this research area can be found in section 3.2.2.

Reflective thinking, considered the key aspect of the self explaining process, is defined as the "active, persistent and careful consideration of any belief or supposed form of knowledge in light of grounds that support it and the further conclusions to which it tends" (Dewey, 1910, p. 4). This type of thought is not exclusive to self explaining and can exist as part of a collaborative process to support information and knowledge exchange. It is this reflective and sense-making aspect of explaining that may be responsible for the beneficial impact on confidence and learning outlined below. A full summary outlining the research into the beneficial impact of explanation and argument on decision confidence, learning performance and facilitation of critical thinking can be seen in Table 1.

2.2.2 Confidence Effects

The enhancement of perceived confidence held in a decision as a result of explanation is of particular interest in research and to this investigation. Confidence in a decision as a function of explanation (or indeed, argument) could be fostered in a variety of ways during the decision making process. As an example, explanation of a decision may help to reveal possibilities that would not be available if the process of explanation had not been undertaken. Koehler (1991) suggested this confidence effect is created by an increase in the depth of processing facilitated by constructing an explanation during a decision making task. This in turn increases belief in the truth of decision from examining the evidence more rigorously and thus confidence in the decision being made.

Table 1 Summary of research into the beneficial impact of explanation and argument on decision confidence, learning performance and facilitation of critical thinking. The possible mechanisms by which explanation and argument can improve these aspects are also detailed in the linked benefits.

Authors	Reasoning Type		Mechanism	Linked Benefit		
	Explanation	Argument		Confidence	Learning Gain	Critical Thinking
Chi et al (1989)	x		Reflection & Inference		x	
Williams et al (2010)	x		Pattern Discovery	x	x	
Koehler (1991)	x		Depth of processing	x		
Lu et al (2011)	x		Justificatory approach	x		
Koriat et al (1980)	x		Belief increase	x		
VanLehn & Jones (1993)	x		Knowledge Gap Filling	x		
Schank (1986)	x		Impasse driven Learning	x	x	
Sieck & Yates (1997)	x	x	Rationale construction	x		
Wolfe (2011)		x	Argument construction	x		
Burge & Brinkman (2010)		x	Rationale Construction			x
Chi, De Leeuw, Chiu, & LaVancher (1994)	x		Reflective thinking			x
Chamberland et al (2011)	x		Reflective thinking		x	x
Osborne et al (2004)		x	Balanced arguments		x	x
Berthold, Eysink, & Renkl (2009)	x		Depth of processing - engagement		x	
Bielaczyc, Pirolli, & Brown (1995)	x		Depth of processing - engagement		x	
Langer and Applebee (1987)	x		Wider attention to material		x	

This suggestion that confidence is facilitated by explanation via increased belief is supported by research conducted by Lu, Chiu, and Law (2011). The findings suggested that a more justificatory approach may foster a perception of confidence. People do appear to prefer providing more complex justifications when generating explanations as opposed to merely stating facts. People may provide a justification as a pre-emptive response to possible disagreement in the future (Kuhn, 2001). This suggests that justificatory explanation may be more akin to argument, if it's purpose it to influence a receiver in some way. The production of a pre-emptive argument may strengthen the author view, thus the apparent preference for justification can result in people perceiving their explanations as more accurate.

This finding also supports earlier research by Koriat, Lichtenstein, and Fischhoff (1980) who confirmed that giving reasons and explanations for a decision may help attenuate overconfidence by increasing overall belief in a choice. The mechanisms which may produce this effect include knowledge gap filling, whereby belief in a decision is increased by elaborating further on principles cited during an explanation (VanLehn & Jones, 1993). In addition, the concept of impasse driven learning (Schank, 1986) may lend insight here, as explaining a decision may uncover inconsistences and force people to find information to overcome this, fostering a sense of achievement.

Further evidence of the confidence bolstering impact of explanation was presented by Sieck and Yates (1997). The work indicated the presence of an explanation-confidence effect. The research encouraged participants to formulate arguments to support their choices using the 'Asian Flu Problem' taken from the original study by Miller and Fagley (1991). The production of a rationale for their choice appeared to increase confidence in the decision. Again, this effect may be due to the complex reasoning involved in producing an explicit argument which may foster a sense of understanding in the task and bolster confidence. Another key aspect of this study is the inclusion of a 'planning' condition, in which participants were instructed to 'think' about an argument they might produce. This condition was no more effective than giving responses alone. This indicates that there is something unique about explicitly constructing an external argument and that this process of producing a coherent, organised argument is crucial to facilitating the observed beneficial effects.

The act of expressing your decision in an external rationale format may help clarify the process you undertook and any trade-offs you encountered, particularly in an unfamiliar decision scenario. Sieck and Yates (1997) also suggested that those in the explanation group

may have attended to the task more and would, if prompted, demonstrate a higher recall of task information than the other groups. This would suggest that confidence is linked to learning and attention. Unfortunately, a post test of task information recall was not performed.

Explanation may also bolster confidence as it performs as a strategy by which patterns in the available evidence can be revealed. These patterns that may provide support for a decision may not otherwise be discovered in the absence of constructing an explanation. This idea was posited by Williams, Lombrozo and Rehder (2010). The research indicated that eliciting an explanation for a decision forces people to create patterns and links between information that may not be immediately obvious. They referred to this effect as a 'subsumptive constraint.'

Williams et al (2010) argue that the subsumptive constraints of explanations are only beneficial if the patterns inferred are in fact correct or exist at all. This is less of a concern in ill-structured domains where patterns can be observed and presented as long as they are satisfactorily justified. In fact, this tendency to seek patterns could be considered a useful side effect if the goal of the activity is constructing coherent and logical explanations. This suggests that rationales may help in sense making of newly presented material if the explanations within sufficiently process the available information.

Increased confidence in a choice as a result of an externalised explanation may be the result of numerous factors. Superficially, it may be a result of the perceived effort exerted, that is, explanation and argument construction may exert a cognitive load and therefore induce a sense of 'effort.' This increased effort may translate into increased confidence and belief that a strong solution has been produced (Sieck & Yates, 1997). One of the most succinct summaries of the confidence effect in explanation and argument is "the processes (that) we use to convince others are also used to convince ourselves" (Wolfe, 2011, p. 92). This may come to light with further examination of the structures present within rationale style arguments that are intended to have a persuasive impact on others and in turn, may act on the confidence held by the author.

2.2.3 Explaining and Learning

Explanation is thought to trigger a number of cognitive mechanisms which give rise to an observed beneficial impact on learning. The use of explanation alone as a learning

intervention has been shown to outperform groups which were controlled for motivation, task processing time and attention (Chi, De Leeuw, Chiu, & LaVancher, 1994). Self explaining has been shown to be particularly useful for facilitating the acquisition of conceptual knowledge (Berthold, Eysink, & Renkl, 2009). The activity of explanation may increase engagement with the learning material and encourage the learners to integrate new information with prior knowledge. This process allows 'inferences' to be generated that can fill in any knowledge gaps. Explanation has also been shown to facilitate skill acquisition by perhaps making elements of task more memorable (Bielaczyc, Pirolli, & Brown, 1995). With these mechanisms in play, it is reasonable to assume that self-explanation does exert a considerable cognitive load, but it is this exertion that increases engagement and deepens processing while self-explaining.

The Levels of processing theory (Craik & Lockhart, 1972) has been proposed as a possible mechanism for how explanation and argument may assist in learning tasks. The theory is not uncontroversial, particularly in light of emerging neurological data for memory storage. However this theory may lend insight into how memories are actually encoded as a process, irrespective of specifying where memories are stored. The theory may offer an explanation of why certain actions and tasks lead to greater learning compared to others, without direct (or observable) rehearsal strategies. Such activities as self explaining or argumentation have been shown to assist in the learning of procedural, conceptual and factual knowledge. There may be evidence within these externalised explanations and arguments that give an indication of the depth of processing of the material presented, in terms of how the material was assessed, analysed and critiqued by the author during construction. So far this intricate approach to analysis has not been effectively carried out.

Evidence for the process of externalising explanations in particular as being superior to other forms of writing in term of learning support has been proposed by Langer and Applebee (1987). The research compared three different types of writing groups and a non-writing group. The comprehension questions and summary group were more successful than no writing; however the analytical writing task outperformed all groups in terms of the retention of particular parts of the text. It appears that the analytical group attended to those parts relevant to their task and processed them more extensively. This could suggest mechanisms for how argumentation can help in retention, as arguers attend to and evaluate evidence relevant to their particular position. The authors proposed this finding was a result of increased manipulation of the material. This lends credence to the notion that the levels of processing theory may be a valid partial explanation for these types of effects.

Further evidence for the unique benefits of generating an explanation is clear from research examining the distinction between generating an explanation and attending to an explanation. It appears that the formulation of a personal explanation triggers a unique process when compared to attending to a pre-constructed explanation. The work of Hausmann and VanLehn (2007) focusses on this distinction. The research compared paraphrasing of existing explanations with the construction of new explanations when learning physics concepts. Even students who paraphrased high quality explanations were outperformed by those who constructed their own explanations, regardless of quality. These results suggest that explanation involves active engagement with materials and that explicitly constructing an explanation is the effective part of the process. This process of generation may involve assessing prior knowledge and its relevance to the current task. Jacoby (1983) suggests that individuals are more likely to recall or recognise items at a later point that they have produced themselves.

2.3 Direction of Argument and Explanation

2.3.1 Introduction

The discussion of the research into explanation and learning would suggest that rationale style arguments may help decision makers to engage more deeply with available resources and make more critically informed choices. These findings may additionally be influenced by the intended direction of the rationale or the actual interactivity such as the presence of a collaborative group.

This section will examine the study of argument and explanation in terms of how the direction can impact learning performance and possible structures within. Argument in particular can be studied in both individual and collaborative contexts and both provide invaluable insight into the process and purpose involved in argument construction.

2.3.2 Persuasion and Argument

One particular aspect that is pertinent when considering rationale direction is that of persuasion. If an argument is intentionally other directed it could be considered to be intended to have a persuasive influence on a receiver. Additionally, if the author holds a view that the argument will be used by others in the future, it would be reasonable to assume that they may wish to influence a reader in some way. The persuasive ‘strength’ of an argument

depends on numerous factors, many of which depend on the properties of the receiver. A rationale, if constructed within a group environment or used to defend a decision in the future, may have a persuasive intention, or at least, be constructed as such by the author. The original definition of rationale style argument encompasses the idea that these arguments could be intended to convince others, intentionally or otherwise. As persuasion is often used as a measure of argument strength, and therefore an aspect of 'quality' it is important to briefly discuss some of the relevant considerations when ascertaining the persuasive property of an argument. Attempts to empirically assess argument quality will be discussed in section 3.2.

The role of the receiver of an argument is a central consideration when examining persuasion as the actual persuasive impact of an argument lies with the influence the argument has upon the receiver, possibly independent of the author intention. Social Judgement Theory (Sherif & Sherif, 1967) separates the position that a receiver may hold relative to a new position into the latitudes of acceptance and rejection. The former includes positions that the receiver of an argument or explanation may find acceptable, the latter, includes the positions that a receiver finds unacceptable relative to their own. In order to be persuasive, the argument needs to fall within the latitude of acceptance for a receiver. This dimension of argument strength, the persuasive effect on the receiver, may be entirely independent of the actual structure or validity of the argument in an objective sense as even a well-structured, balanced and valid argument will be unacceptance and thus not persuasive if the position it is arguing for falls into the latitude of rejection.

A second aspect of persuasion is that of receiver involvement (Johnson & Eagly, 1989). This aspect is usually divided into three distinct types. Firstly, value-relevant involvement refers to the values or beliefs that a receiver holds that may impact where the latitude of acceptance and rejection fall (e.g. a Green Party member may have a wider latitude of rejection for urban development). Secondly, outcome-relevant involvement refers to specific outcomes from a position that has a personal relevance to the receivers (e.g. people who have been diagnosed with a terminal illness may have more involvement with legislation on end of life care). Finally, the third aspect, impression-relevant, refers to the receiver's concern about the impact their position may have on others (e.g. a lecturer may worry about the views of their colleagues). The level of persuasive power that a message may hold is dependent on these complex factors which are difficult to predict due to individual differences.

This thesis will aim to ascertain whether an author intends their argument to be persuasive and whether they feel it may have a persuasive impact on others. As the explicit investigation of the actual persuasive impact of the rationales is outside the scope of this thesis, it is primarily the author's perception of the persuasive power of their arguments, as an additional indicator of perceived confidence in the rationale itself that will be considered.

A popular method of examining the impact of the direction of argument is to juxtapose different levels of interactivity in the hope of highlighting differences in response to these. One of the most common comparisons evident in the literature is self-directed versus jointly constructed explanations. General argument models are useful as far as determining which structures may be present within an argument, but the perception of direction may be significant in determining the prevalence of certain structures over others. This has not yet been fully explored. Arguments that are either self or other directed appear to warrant different considerations. Some of these will be discussed in this section along with the apparent influence that a self, joint or other directed argument has on task performance and attitude.

2.3.3 Other Directed Explanation and Argument

As the benefits of purely self explaining have been previously discussed, the apparent further benefits offered using collaborative activities and other directed explaining on task performance will be briefly presented in this section. While the research discussed here does not explicitly address the issue of the perception of direction held by the author of an argument, the findings are used to help understand and consider the implications of this perception and inform the methodology of future research into this area.

To begin to fully understand the potential impact of the perception of direction, the differences between self and other directed explanation need to be fully defined. Hausmann, Chi and Roy (2004) hypothesised that any learning benefits observed from collaborative work are a result of these three mechanisms:

1. Other directed explanations – explaining to another
2. Co-construction – explaining with another
3. Self-directed explanations – explaining to oneself

It would appear that all three types of activity can occur in a collaborative context, the question is whether other directed explanation has in fact occurred in a supposedly self directed context. It is often difficult to ascertain which of these three mechanisms are responsible for differences in explanation quality or learning outcomes. It could be argued that any explanation externalised in the presence of another may in fact be considered other directed. One of the ways to differentiate between them would be to ask participants to self-report the intended recipient of their explanations, a strategy which has not yet been implemented in the explanation research.

Vygotsky (1978) originally suggested that students are capable of performing at higher intellectual levels when working in groups than working alone. This may be due to increased opportunities for explanation and argument which students are less likely to implement privately to themselves. Early research by Heath and Gonzalez (1995) suggested that other directed explanation could yield useful outcomes. The participants reported increases in confidence after explaining their predictions and judgements to others. The increases in confidence post explanation were not justified by increased accuracy alone but appear to be a result of interaction. Interaction may be offering another process beyond information sharing which in turn increases confidence in the decision. This process may be the opportunity to construct a 'rationale' when interacting with another person. This rationale construction could be the strengthening factor in the confidence of the individual decision.

More recently, Cooper, Cox Jr, Nammouz, Case and Stevens (2008) investigated the effect of collaborative groups on problem solving strategies. They concluded that students in interactive groups are forced to become more thoughtful about their actions, that is, it increases metacognitive strategies and justifications for decisions. Again, the results and improved strategies relate to the three mechanisms of explanation listed above. Collaborative groups give individuals the opportunity for self-explanation and also other directed explanations. This could be the cause of the loss of clarity in research as it is not entirely appropriate to directly compare self-explanation with joint or other directed explanation as self-explanation also occurs in interactive situations. Similar research carried out by Hausmann, van de Sande and VanLehn (2008) found that working in pairs increases social accountability, therefore students are more likely to choose better strategies. Again, no further investigation of the content of the explanations was carried out. The self-explanation effect is considered reasonably robust, therefore the indication that the effects can be further enhanced by other directed explanation is an area worthy of examination.

A less direct examination of the impact of collaborative contexts on argument was conducted by Lin, Hong and Lawrenz (2012a). The research compared online group construction of arguments with a pencil and paper condition. Both conditions constructed their arguments as part of a group; however the asynchronous online condition were able to alter their arguments throughout the task. The online group constructed arguments with more rebuttals than the pencil and paper group. The online group also slightly outperformed the paper group in terms of argument quality, defined as arguments which contain more rebuttals against other arguments. This research utilised a quality scale developed by Osborne, et al. (2004a), and which will be discussed further in section 3.2.3.4. All students had access to the arguments of their peers while constructing their own arguments online, which may give rise to an increase in rebuttals. This may be a result of responding to arguments presented by others, or it may also be a function of perceiving their own arguments as being subject to increased scrutiny by others and thus adopting a better argumentative strategy.

This impact of interaction on decision making is known as the 'interaction hypothesis.' This effect was examined by Hausmann, et al (2004). The hypothesis stated that joint explanation would lead to better problem solving performance as there are more opportunities to be interactive. The assistive scores for the problem solving groups, which comprised of the total number of hints used plus the errors made, showed that a joint explanation group used less hints and made fewer errors. Joint explanation appears to enhance the effectiveness of the self-explanation effect above self-explanation alone.

Joint explanation may provide a social cue to avoid glossing over material so explanations are of a higher quality than those in self-explanation. Although the explanations were not qualitatively coded to uncover if there were any differences in the structural content between self or other directed explanation, they were categorised according to whether they were self-directed (reflective) or purposefully other directed. Other directed explanations appeared to be the critical aspect in terms of assisting learning in half of the cases. Self-explanation was also effective for the individual generating the explanation (79% gain from pre test scores), with only a marginal benefit of 29% gain observed for the listener. This finding is predicted by the 'Content versus Generation' hypothesis of explanation. If the content of an explanation was as effective as the production, the benefits would be comparable for both groups. All three mechanisms appear to help learning but to various degrees. Again, this research neglected to perform a full analysis on the content and

structural differences between the self or other directed explanations, but made assumptions of explanation quality based on task performance.

It is unclear whether the additional cognitive load imposed by interacting with others and producing other directed explanations actually impedes or improves learning in this context. The perception of others or a future use for an explanation may also impact on explanation structure and content. As suggested by Ploetzner, Dillenbourg, Preier and Traum (1999), we might adapt our explanation when the listener is merely 'imagined' or not currently present as humans have a tendency to frequently construct only partial explanations if no one needs to read or understand the argument. Regli, Hu, Atwood and Sun (2000) suggested this may be due to the explainer making assumptions about the future uses of their explanation and potential reader background and knowledge.

Much research has focussed on the comparison of self reflective and fully interactive arguments and has posited inferences about the benefits of the two approaches. However, the research into self-explanation has shown how audience or perhaps, perceived audience could impact upon argument quality, by eliciting more other directed explanations. It has become apparent that the adoption of a self or other directed argumentative approach is influenced by a perception of direction and argument purpose. The research has unintentionally muddied the waters in some respects, as self directed activities may still involve a perception of arguing for another and collaborative arguments may also incorporate self directed, reflective activities.

2.3.4 Methodological Issues

In the previously discussed literature for the self explanation effect, the methodology intended to elicit self directed as opposed to other directed explanations. However, closer examination of the research methods reveals that the distinction between explaining for oneself in a reflective manner or for others in an argumentative manner may be misconstrued. Very little research currently exists that explicitly examines the structural and content differences which may exist between perceived self-directed and perceived other directed explanations. There is a lack of understanding of the structural differences that may exist and whether the differences in perception, as opposed to actual interaction, may help to explain any observed differences in learning performance or overall confidence.

Ploetzner et al (1999) proposed that there are five levels of interactivity which can be studied when examining the distinction between explaining for self and others:

1. Explaining to oneself – no listeners and no sharing of explanations
2. Explaining to a passive listener – who is unknown by the explainer
3. Explaining to a passive listener - who is known by the explainer.
4. Explaining to someone who responds in a constrained way.
5. Mutual explanation - explaining to each other freely.

As can be seen in these distinctions, there may be opportunities for further interactivity within each boundary. Research has indicated that self and other directed explaining occurs in both self and other directed contexts (Hausmann, et al., 2004). Investigations into how explanations are constructed with less interaction may help to reveal the processes which might take place in a fully interactive environment. It is important to place constraints on interaction levels in terms of research, in order to be able to infer predictors on behaviour. These levels of interactivity need to be studied for their inherent differences and constraints on task performance, attitude and the externalisation of arguments.

On the far end of the spectrum is explaining privately to oneself, considered a reflective process. These explanations may appear, in comparison to other directed explanations, to be less formal, less complete, less coherent and therefore qualitatively different from explanations directed to a listener. However, these self directed explanations still need to be expressed if they can be empirically studied, which can add a difficulty if there is a sense of presence when these explanations are elicited. Hence, interference in the interactivity levels may occur. This makes fully understanding and bounding the impact of context on argument structures problematic.

As self-explanations are more concerned with repairing the explainers' own mental models, it has been suggested that they should be more powerful than other directed explanations in terms of a beneficial impact on learning. However, the original research into self-explanation (Chi, Bassok, Lewis, Reimann, & Glaser, 1989a) prompted learners to explain aloud to an experimenter, and if Chi et al's definition is to be accepted, this original research was evidently not demonstrating self-explanation but in fact other directed explanation. Similarly, in the work by Chamberland et al (2011) that demonstrated the effectiveness of self-explanations, the explanations were elicited via a think aloud method, which means they may have been perceived as subject to scrutiny by others. Additionally, those in the silent group may have also generated self-explanations in their heads, independent of the conditions. Explanations that are directed to oneself but listened to by an experimenter are

possibly constructed more carefully than explanations directed to peers because of the perceived status of the experimenter.

Similar methodological issues are seen in research by Rosé, Bhembe, Siler, Srivastava and VanLehn (2003), Hausmann & VanLehn (2007) and Renkl (1997) in which participant explanations were elicited by talk aloud protocols, yet considered in the analyses as self-directed. It could be legitimately argued that the explanations were in fact directed at the experimenter and therefore were not self explanations in the defined sense. This tendency may present a case for misidentification of explanation in the literature, as explanations labelled as self directed may indeed be other directed. This presents an apparent intertwining of the self and other directed explanation effects.

This has a potential impact on determining the quality of results. If the perception that the participants hold of whom they are explaining for vary widely within or between groups this may result in apparently inexplicable variations in argument quality. To remedy this it would be useful to consider interaction on a more closely related spectrum, as subtle changes to interactivity can be implemented. This would enable the focus to be on the perception that the participants hold of who they are generating an explanation for. This focus may uncover whether it is the perception of interactivity, rather than physical interactivity itself that may impact on the quality of arguments, explanations, and task performance.

2.3.5 Summary

A distinction can be drawn between external constraints of the audience and an internal constraint of the perceived direction (either self or other). In terms of reconciling the effects of self versus other directed explanation, it could be argued that it is the perception held by the author of the level of interactivity that is in fact the most influential factor, perhaps independent of the actual physical context such as the proximity of others. It is this internal perception of the intended direction of an argument, and its possible impact on argument structure and decision making that will be a central theme in this thesis.

The research discussed so far has informed how rationale style arguments may be beneficial for individuals and groups in terms of task performance and decision confidence. The next chapter will outline some of the most popular methods for argument content and structure analysis and the varying approaches for evaluating argument and explanation quality.

3 Argument and Explanation: Content and Quality Analysis

3.1 Structural Content Analysis

Prior to outlining some of the available methods for assessing argument quality in Section 3.2, the basic approaches for examining argument structure and content will be addressed here. Many of these content analysis approaches form the basis for quality analysis frameworks that will be discussed in section 3.2.3.

3.1.1 Introduction

Models of argument have been proposed to serve several purposes in various research disciplines and educational approaches. These approaches are more systematic and informative than the typical broad categorisation approach that is often adopted when analysing arguments in research. Three of the most typical uses for argument models are analytical, normative and descriptive (Nussbaum, 2011). Analytical models, the most common types of argument frameworks, allow researchers to breakdown arguments into components to reveal coherence and the overall structure of an argument. Models are also used in normative ways, to judge the strength and quality of a particular piece of text or the components. These types of frameworks are less common as they usually rely on domain knowledge. Finally, the least common types of frameworks are descriptive models. These can be used to make explanatory claims about 'how' people tend to argue. No particular model of argument can encompass all three approaches, and the research objectives will ultimately determine the utility of any approach.

3.1.2 Toulmin Model of Argument

The Toulmin (1958) model is one of the most popular frameworks for studying argument structure. Perhaps due to its intuitive component structure and ease of implementation into research methodology, the Toulmin model has a wide and varied use in argumentation research. The framework comprises of a list of six components listed below, which may indicate argumentative structure:

1. Claim: the position or claim being argued for and the conclusion of the argument.
2. Grounds: reasons or supporting evidence that increase belief in the claim.

3. Warrant: the principle or chain of reasoning that connects the grounds to the claim.
4. Backing: support, justification and reasons for the argument
5. Rebuttal or Reservation: concessions or flaws in the claim and counter-arguments.
6. Qualification: specification of limits or scope to claim, warrant and backing.

The Toulmin model does propose a slightly more fine grained examination of an argument in the form of 'qualifiers' (also known as 'modal qualifiers'). These qualification elements are words and phrases which indicate the strength of a claim or evidence item, such as 'mostly' and 'definitely'. The qualifier element of the argument may indicate how strong the arguer feels a claim may be, but is not an indication of the objective strength of that claim for a receiver or whether the use of the qualifier is in fact valid or accurate.

Toulmin (1958) took care to warn that this framework did not comprise a descriptive 'theory' of argumentation to inform the notion of 'how' people argue, but rather, a way to examine structural elements that may exist within an argument (Van Eemeren, Grootendorst, Johnson, Plantin, & Willard, 2013). In addition, a common misconception of the Toulmin model is that an argument must contain all six elements, which again, is not the case; in fact many elements (particularly warrants) may be implicit or entirely absent (Stein & Miller, 1993). This model does not attempt to apply a deterministic view on the quality of an argument, this remains domain specific and is determined by the use of available evidence and its perceived strength within that domain. In this regard the model is not descriptive of actual psychological processes. In spite of this the model has been adapted as a measure of argument quality and this will be discussed in section 3.2.3.4.

3.1.3 Rhetorical Structure Theory

3.1.3.1 Background

The original Rhetorical Structure Theory (RST) framework was developed by Mann and Thompson (1988) as a framework for analysing text coherence and structure in written monologues. RST has gained coverage in recent years for its focus on fine grained linguistic features and markers within text, which can potentially be identified by automated parsers. RST helps to identify un-signalled discourse relations by examining language structure, the intended effect on the reader and explicit discourse markers. A full list of relations, including the original definitions and some more recent additions are listed in Appendix 2.

3.1.3.2 Analysis Procedure

The process of rhetorically analysing a piece of text begins with determining the individual units of text which are known as elementary discourse units (EDU). Once this has been established, the relations between each unit can be identified. An example of the EDUs contained within a Condition and Concession relation can be seen in Figure 1. When looking at two EDUs there is more often one unit which is considered more vital to the text coherence, this is labelled the nucleus. For example the nucleus element for the Concession relation in Figure 1 contains the main point of the statement, with the satellite offering additional information. The nucleus is identified on the basis that if this element were removed, the text would no longer make sense. The EDU of lesser importance is labelled the 'satellite.' The satellite is usually a 'supporting' element that may add additional power or plausibility to the nucleus. Once the satellite and nucleus have been identified, the relationship between the two can be determined. Once the analysis of a piece of text is completed, a distinctive 'tree form' is often produced (see Figure 2). This diagram represents how the text elements are linked to each other to form a coherent argument or statement.

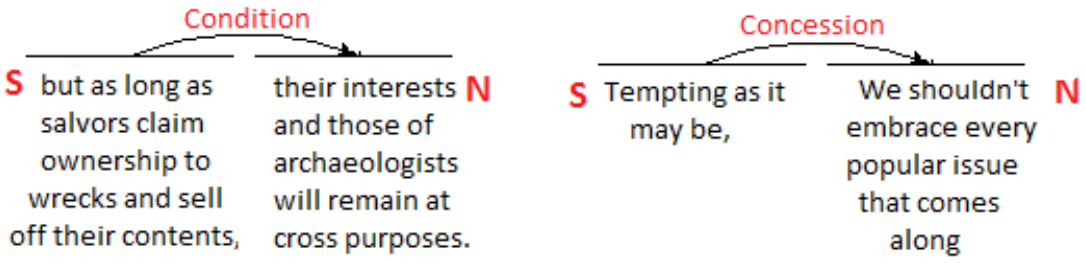

Figure 1 Tree diagram RST relations of Condition and Concession showing Satellite and Nucleus units. Taken from Mann and Taboada (2015).

The rhetorical relations are also split into categories, depending on the assumed intention of the writer; presentational or subject related. Presentational relations are those which prompt an inclination in the reader, such as an increase in the level of belief in the nucleus (claim), these include providing evidence for a claim or a justification. Subject matter relations are those which explicitly signal the relation in question, these include a Condition, which is signalled by 'if' and a Means relation which may be signalled with 'by'. These are not intended to prompt any inclination or positive regard in the reader.

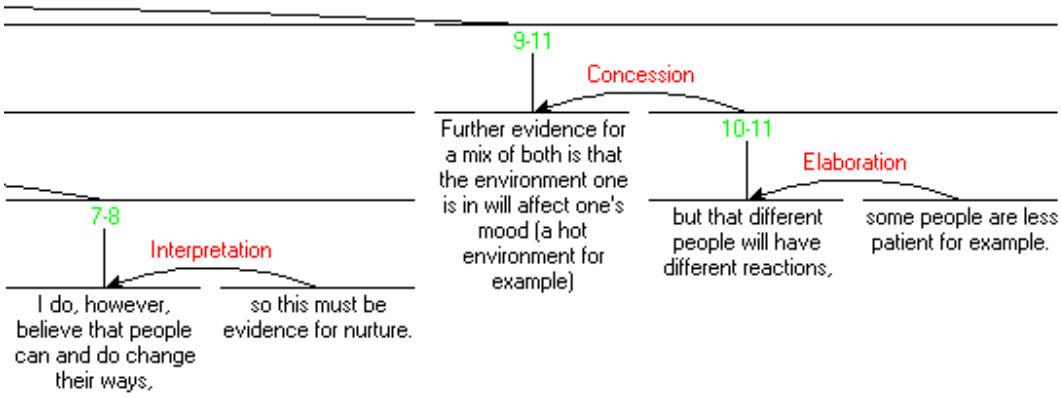

Figure 2 Example of partial Classical RST tree analysis

Mann and Thompson recognise that the analysis process of determining the intention of the writer relies almost solely on individual judgement. When a relation has been identified the outcome is referred to as being "plausible to the observer that it was plausible to the author writing the text, that <the finding> holds." (Mann, 1999, p.4). The authors suggest that if an analyst discovers a text span for which both a subject matter and presentational relation are applicable, then the subject matter relation should take precedence. This does pose a difficulty when analysing argumentative text, as the purpose of the text is predominately to have an impact upon the reader. Thus it would be expected that presentational relations are more informative in this context. Rater and annotator agreement will almost always vary as a result of individual differences, as the impact on the reader is central in argument analysis, therefore the individual perspective will influence the extent and perception of this impact.

One of the fundamental difficulties in using RST is that the relation definitions are sometimes ambiguous as is the text being analysed. This inherent property of applying RST is known as multiplicity. This is due to the identification of relations hinging upon plausibility judgements. This can lead to the possibility of a text being legitimately analysed in several ways. An example of this multiple analysis possibility can be seen in a letter extract taken from Mann and Thompson (1992) in Figure 3.

Segment 11 ZPG's 1985 Urban Stress Test, created after months of persistent and exhaustive research, is the nation's first survey of how population-linked pressures affect U.S. cities.

Segment 12 It ranks 184 urban areas on 11 different criteria ranging from crowding and birth rates to air quality and toxic wastes.

Segment 13 The Urban Stress Test translates complex, technical data in an easy-to-use action tool for concerned citizens, elected officials and opinion leaders.

Segment 14 But to use it well, we urgently need your help.

Figure 3 Letter extract to illustrate a possible alternative Rhetorical Structure Theory analysis

The segments 11 through to 14 were labelled as a Concession relation, with segment 14 highlighting a flaw in the use of the Urban Stress Test, and is explicitly signalled with the use of a 'but' discourse marker. However, in terms of interpreting the intention of the author, another analyst may believe that it is plausible that segments 11-13 offer information that enable more effective understanding of the next segments and thus provide a Background relation function, over and above the apparent Concession relation. This possible variation of judgment is considered normal in this type of linguistic analysis as it is the impact of text on the receiver that may be variable depending on the individual analyst.

Often there is no way to reconcile these differences as even when using several analysts, coming to an agreement can be troublesome. This multiplicity may result from a lack of required discourse markers in natural text and is most often due to differing plausibility judgements by analysts. Taboada (2006) suggested that 60%-70% of markers may be implicit in the text (see section 3.1.5 for an example of an 'implicit' relation), and need to be inferred by examining adjacent EDUs. This may be why a 'handbook' of discourse markers to assist in the reliable identification of relations with text is not forthcoming. These issues could be addressed more effectively with more extensive use of RST and comparisons and integration with other models to strengthen the inferences made from the rhetorical analyses. The infrequency of common discourse markers within text along with the multiplicity inherent in RST poses difficulties for any attempts to construct automated analysis techniques based on RST. This issue will be discussed along with an examination of the available RST and argument based natural text analysers in the next section.

There is no agreed theory of language that could be considered the basis for RST. It offers a framework of possible structures with which to build a coherent deconstruction of the text. There is also no indication of the power or weight of certain structures in terms of persuasion or argument quality as a result of the lack of theoretical grounding. RST can be used as a data gathering tool to identify structures and patterns in texts and for raising questions about the function of monologues and may prove useful for suggesting in part, what language models for certain contexts might look like via repeated analysis of a corpus (Mann & Thompson, 2002).

3.1.3.3 Practical Research Using RST

Research utilising the RST framework has demonstrated that it can be used effectively for analysing argument structure (Azar, 1999; Green, 2010). In RST the distinction between the two key elements of a piece of text, the satellite and the nucleus, is considered a typical argumentative structure not dissimilar to the Toulmin model i.e. a claim and its backing. RST was considered in Azar's (1999) paper to have five relations specifically orientated to presenting arguments, namely: Evidence, Motivation, Justification, Antithesis and Concession. The full definitions of these relations are available in Appendix 2. Each relation was posited as a type of argument in itself, not unlike Walton's argument schemes. For example, the Evidence relation pertains to a supportive argument, Motivation to an incentive argument, the Justify relation to a justifying argument and the Concession and Antithesis relations being intended as persuaders. However, caution must be taken when attempting to use RST to give any insights into the cognitive aspects of argument, as it is not such a theory, simply a framework of labels with which to identify the externalised text or utterances into structures of a rhetorical nature with an intentional basis.

Mentis, Bach, Hoffman, Rosson and Carroll (2009) utilised the RST framework more fully, to investigate how the style of rationales might change through the course of a collaborative activity. During the analysis, 12 rhetorical structures (from the possible set of 32 Classical RST relations) were identified in the corpus. These were grouped into categories of 'State' to indicate information presentation, 'Argue' to indicate an argument for or against an idea and 'Analyse' to indicate an interpretation of information. This categorisation bears a resemblance to the Toulmin model of argument, as State relations could function as Backing and Argue could be considered as performing the same function as a Rebuttal. In the beginning of the activity the group member utterances comprised mostly of statement type

relations such as Conjunctions Restatements and Elaborations. These relations, according to the original RST definitions pertain more to statements of fact and knowledge rather than argumentative type text intended to portray or convince a receiver of a viewpoint. Towards the end of the interaction the participants had come to rely upon Evidence and Antithesis relations more heavily. These relations are intended to argue the case for a view and decrease agreement with the opposing view, both by acknowledgement of flaws in the main claim or by discrediting the alternative position. The observation that Argue type relations appeared to increase over time suggests that participants initially examine information and then analyse and determine its merits. However, the State category of relations was still prominent throughout the interaction indicating that knowledge sharing continued throughout the problem solving process. The methods of categorising rhetorical relations in terms of argument purpose is often done post-hoc but without empirical backing for the assignment. This thesis will endeavour to categorise rhetorical relations into roles that denote argument purpose based on the empirical findings.

More recent work has been conducted that demonstrates support for these findings and validation for the use of RST in research methodology. Xiao (2013a) examined the differences in reasoning styles that may exist between teams working in a shared virtual workspace. Participants initially focussed on using subject matter relations. These first rationales were primarily comprised of Circumstance, Elaboration and Evaluation relations. These relations are used to offer contextual and background information. An interesting finding is that five out of the seven relations that were not found in the corpus were concerned with the persuasion of others. This may be a result of subject matter relations being more appropriate for self-reflective commentary, instead of more direct dialogue style interaction which would require a persuasive approach. The participants did appear to use more of the presentational category of relations, concerned with persuasion, over the course of the task. This may be a result of a shift from reflective thinking to communicating persuasively with others.

In essence, if RST is to be used to fully analyse arguments or examine the importance of structures, it needs to be used in a holistic way to give a richer insight, as opposed to focussing on one particular relation, however this is possibly easier to conduct. In spite of the laborious and subjective application process, RST has been successfully incorporated into research methodology and has highlighted some interesting trends in reasoning styles that may not have been revealed using other argument analysis approaches.

3.1.4 The HILDA parser

The "HIgh-Level Discourse Analyser" (HILDA) text parser was developed by the Global Lab project as an automated tool that is accessible via a simple online interface (Hirst & hernault, 2015) to identify EDUs within a text and assign relations based on RST (Feng & Hirst, 2012; Hernault, Prendinger, & Ishizuka, 2010).

The HILDA parser is based on the RST style corpus analyses conducted by Carlson, Marcu and Okurowski (2003) who created a set of 78 relations (53 mononuclear and 25 multinuclear) which were organised into 16 categories. These categories were supposed to contain the relations that shared rhetorical meaning to a certain extent. These higher level classes are the labels assigned by the automated HILDA parser. A full list of the classes and the sub level relations can be found in Appendix 3. As the parser will only identify the higher level class and not the more specific rhetorical relations within the class, the analysis is fairly broad in this respect.

The tool also highlights 'attribution' aspects using markers such as 'I' and 'Me.' The Attribution label is not a rhetorical relation. It captures ownership between agents (or the author) and abstract objects within the argument. The HILDA parser seems to favour assigning more subject matter relations to EDUs. The Same-unit relation is used to link spans that are separated in part by an embedded unit or span (Carlson, et al., 2003).

The HILDA parser requires that each text be uploaded individually and it is then automatically segmented by the tool to identify sentences and paragraphs within the text. A sentence is tagged with <s> and the end of the rationale is tagged with <p>. The output is presented in a tree form with each text section tagged with the appropriate relation. An example of the output from the HILDA parser is shown in Figure 4 . An attempt is made to assign satellite (S) and nucleus (N) labels to the text spans.

Attribution [S,N]
I believe
»Elaboration [N,S]
that people are born with an innate level of `aggression'
»Elaboration [N,S]
that is influenced by genetics and hormones such as testosterone, and that the level of aggression seen in an individual can be managed through learned behaviour.

Figure 4 Example of HILDA parser output. (N-Nucleus, S-Satellite)

HILDA represented an important step in discourse analysis as previous systems such as SPADE (Soricut & Marcu, 2003) were limited to sentence level analyses, whereas HILDA is finer grained and can process all types of text. In trials the HILDA parser reached 78.3% agreement with human analysts for the labelling and analysis. The primary application of the parser thus far is that of dialogue generation from standalone text. The argumentative elements are identified and labelled by the text classifiers and can then be converted into an argumentative dialogue, this step is done after a rhetorical analysis of the text. This system has the potential to be applied to an educational text analysis as it is domain independent, fast, reliable and easy to access. However, the parser still requires knowledge of RST to reliably comprehend the output and a diagnostic element of feedback to assist in argument quality assessment is lacking.

The intentional nature of any argumentative text is probably only discernible from the broader context of the text and therefore more sensitive to a manual analysis. This may be an intrinsic factor of any automated approach that relies solely on determining relationships based on explicit discourse markers, and not the imagined intention of effect of the writer on the reader.

It is important to note that while these automated RST based parsers do not always perform to the standards expected, it is not a reflection on the utility of RST in analysing arguments, but rather that there is still a gap between the understanding of the applicability of RST and the needs of machine language on which the tool is built. Parsers have since been developed that are claimed to be superior to HILDA in terms of identification of relations, as they are able to incorporate both explicit and implicit discourse markers. However these parsers may not have been fully developed as they are not yet available for use or research purposes, meaning the claims cannot be verified.

3.1.5 The Penn Discourse Tree Bank Parser

A possible alternative to the HILDA parser which analyses text to a similar degree of granularity is the Penn Discourse Tree Bank (PDTB) parser. The parser is also available to access as an online tool (Lin, Ng, & Kan, 2015) and was developed by Lin, Ng and Kan (2012b) to demonstrate and utilise the discourse tagging style used to analyse the Penn Discourse Tree Bank corpus. The parser was trained to process any text by initially identifying all discourse relations and label the arguments present in the text. The parser was trained by analysing a large machine readable annotated corpus.

One of the most unique features of this parser, compared to others, is the attempt to signal not only explicitly signalled relations, but also implicitly signalled relations. The implicit relations are not signalled directly in the text and usually require a human analyst to identify their presence. The following example is an implicit Contrast relation taken from Lin, Ng and Kan (2012).

"She was untrained and, in one botched job killed a client. Her remorse was shallow and brief."

In this example an analyst could infer that the connective word "however" can be inserted in between the two sentences. This would enable the Contrast relation to be implied. In comparison, an explicit relation is signalled by the presence of a discourse marker connecting the two elements. An example of this would be:

"The Treasury said the U.S. will default on Nov. 9 if Congress doesn't act by then."

The Condition relation is clearly signalled by the discourse marker 'if' in-between the two elements (Lin, et al., 2012a). If no relation can be inferred implicitly and no explicit marker is included in the text, a 'non entity' (or 'No Rel') label is applied.

The parser analysis is reminiscent of RST in that it involves segmenting the text automatically into EDUs and then determining the relations that hold. The relations used in the PDTB are similar to but not strictly based upon the original RST framework. The parser does not assign satellite or nucleus labels to text spans. The relations that are most

commonly identified by the parser appear to rely heavily on explicit discourse markers. A full list of the relational classes used in the parser can be found in Appendix 4.

The parser does tend to overlook presentational style relations; that is, those relations with intentional properties, such as Concession in favour of subject matter relations. This is possibly because subject matter relations are more often explicitly signalled in the text. In comparison, in a manual discourse analysis, the marker 'but' can be an indication to the analyst of a Concession to an earlier claim rather than a Contrast. The context and intention are crucial when deciding which relation may hold. An example of the output produced by the PDTB parser is shown in Figure 5.

The PDTB parser accepts text without any segmentation and text can be loaded directly into the interface. The output takes the form of a table, which splits the text initially into arguments. Each argument span has a relation attributed to it, and the connecting words that denote the relations such as 'and' and 'but' are highlighted in red for each argument. The prefix of 'Exp' on a relation label denotes that the relation is explicitly signalled in the text. The prefix of 'NonExp' indicates that the relation is implicit in the text.

	Text
Exp 0 Concession	I think that aggression mainly results from child rearing practices and one's personal experience through one's childhood and teenage years.
NonExp 0 Conjunction	Although hormones and genes might give different people, different pre-conditions.

Figure 5 Example of PDTB parser output.

This parser may help to provide an approach to argument analysis and feedback that is not domain specific, but focussed on natural language coherence, which could be applicable to any context. There is a need for a model of typical 'argument' structure in terms of the rhetorical relations it contains and how these relate to argument quality in order to begin to visualise a generic argument support and feedback system. The automated RST and language parsers may offer a possible foundation for this, as the build and source code for the PDTB and HILDA parsers are freely available to download and adapt.

This parser may have applications for educational argument analysis, however in its current form the output is perhaps not informative for a novice user, without a comprehensive background in linguistic analysis. Similarly, there is no standard model of the expected rhetorical structures that may be present in a good quality argument, so this parser would have some way to go before it could be considered a diagnostic tool for informative feedback on argument quality.

3.2 Quality Analysis in Explanation and Argument

3.2.1 Introduction

The following section will describe some of the attempts that have been made in previous research to assess argument and explanation quality and how these assessments provide further insight into the mechanisms behind the beneficial effects of eliciting explanations that have been discussed above. The following discussion of argument quality will be structured along two dimensions:

1. Section 3.2.2 will discuss research that examines indicators of quality that relate to task performance, such as learning or confidence.
2. Section 3.2.3 will discuss indicators of quality that are structure related and domain independent

Effective explanations and arguments have been shown to be beneficial in terms of decision confidence and declarative knowledge acquisition linked to critical thinking. The structures within and features of the explanations that may be responsible and thus considered to be quality indicators will be discussed in this section.

The definition of quality in terms of argument may depend on the goal of the activity in which it is situated. Scientific arguments may require robust interpretation of available data and evidence (Osborne, Erduran, & Simon, 2004b), whereas arguments within a group context may seek to persuade and thus quality will pertain to the persuasive impact of the argument, possibly including the level of rebuttal of opposing views. Quality may also pertain to those structures within an argument that lead to the greatest gains in terms of task performance such as information retention or argument skill acquisition. In this case

structures such as rebuttals, which are considered to require more skill to produce, may be taken as quality indicators (Kuhn, 1991).

3.2.2 Argument Quality Indictors and Task Performance

3.2.2.1 Introduction

Particular features and types of explanation that have been indicated as important for supporting learning in terms of information recall will be discussed here. Much of the research discussed implies that quality is determined by the presence of structures which correlate with desired task outcomes. A summary of research discussed, detailing the quality indicators and the associated benefits can be seen in Table 2.

Authors	Quality Indictor	Linked Benefit
Berthold et al. (2009).	Types of explanation: Principle and rationale based	Rationale increases conceptual and procedural knowledge
Chi, Bassok, Lewis, Reimann, & Glaser (1989b)	Amount of explanation	Increase in learning gains (post test-pre test)
Chi, et al (1989a)	Types of explanation: Principle and rationale based	Rationale increases conceptual understanding
Ferretti, MacArthur and Dowdy (2000)	Frequency of Toulmin elements as complexity	Improved argument skill – critical thinking
Ferretti, MacArthur and Dowdy (2000)	Persuasiveness rubric – elaborated argument	Improved argument skill – critical thinking
Lin and Lehman (1999)	Style of explanations	Metacognitive explanations facilitate understanding
Lin, Newby, Glenn and Lafayette (1994)	Evaluative explanations	Increase in learning gains (post test-pre test)
Rosé, et al (2003)	Amount of explanation – why questions prompt	Increase in learning gains (post test-pre test)
Rottman and Keil (2011)	Elaborations	Improved argument skill
Sandoval and Millwood (2005)	Use of rhetoric in data interpretation	Improved argument skill
Schworm and Renkl (2007)	Amount of elaborations	Increase in learning gains (post test-pre test)
Von Aufschnaiter, Erduran, Osborne and Simon (2008)	Use of warrants - Toulmin	Improved argument skill
Wolfe & Goldman (2005)	Use of elaborations or paraphrasing	Elaborations enhanced understanding

Table 2 Summary table of argument and explanation quality indicators and the linked benefits

3.2.2.2 Argument and Explanation Length

Broad examinations of explanation structure in previous research have found that the amount of explanation generated is a possible predictor of task performance. This was originally proposed by Chi, Bassok, Lewis, Reimann, & Glaser (1989b) who found that 'good' explainers showed greater understanding of the learning material, as well as better monitoring of their own undertsanding. Poor explainers who generate less content may be more likely to be unaware of inconsistences or gaps in their knowledge.

More recently, Rosé, et al (2003) broadly analysed a large corpus of human tutoring dialogues to tease out the features which correlated significantly with learning gains. The average turn length of each student appeared to correlate with learning; however the total number of words uttered did not. Open and 'why' questions appeared to encourage the students to say more. This research did not examine some of the specific mechanisms relating to the self-explanation effect that may be relevant here, such as evidence of reflective or evaluative strategies, only that turn length appeared to be a key consideration. It may well be that the increased length of an explanation may indicate that it contains a higher frequency of complex argumentative elements and it is these that are responsible for the increased learning gains.

3.2.2.3 Categories of Explanation

Extending the idea that argument length represents a feature of quality, in terms of longer length explanations being linked to increase learning, some research has focussed on labelling specific types of explanations that can be identified. Explanations have been evaluated by assigning categories loosely based on the perceived function of the explanation and the achieved goal. Lin, Newby, Glenn and Lafayette (1994) found that evaluative explanations (referred to as metacognitive) led to the greatest increases from pre to post-test scores. Similarly, Lin and Lehman (1999) found that prompts that encourage participants to explain 'why' they have chosen a particular solution to be more effective in facilitating understanding than prompting for explanations that detail rules or feelings about a choice. These 'metacognitive' styles of explanation often require justification and reflection, which are more argumentative strategies. This increase in complexity may account for the beneficial effects on understanding and retention.

Similarly, Chi, et al (1989a) initially identified a number of broad categories describing 'types' of explanations using a content analysis approach. These categories were later refined and studied by Roy and Chi (2005). These categories were an initial step in

suggesting that certain types of explanations may be of better quality than others in terms of the extent to which they foster deeper learning. The first category referred to principle based explanations. In these types of explanation the learner assigned meaning to a solution step. For example, if ordering was an important step in the problem solution the learner needed to state the following:

''The order is relevant because it does matter in which order you type in the numbers of a PIN"

These types of explanation showed that learners had assigned meaning to a principle by elaborating upon it, either with prior or new knowledge. The second category; rationale based explanation, referred to explanations about the rationale behind the use of a principle in the material. Therefore the rationale-based self-explanations exceeded principle based explanations as they offered reasons for why the principle is as it is, rather than simply stating why it is important in the solution. An example of a rationale based explanation in response to the question: 'Why do you calculate the total acceptable outcomes by multiplying?'' would be:

''Because for the denominator there are five times four branches. Thus, each of the first five branches of the tree diagram forks out in four further branches because each of the first five events can occur in combination with one of the four remaining events.''

These types of explanations showed that learners extended the principles by providing rationales for their understanding and use. This type of explanation exceeds principle based explanation in terms of quality as it provides deeper reasons for the principles being explained. A rationale based approach would therefore be more beneficial in fostering a deeper conceptual understanding (and theoretically, information retention), as procedural understanding is less demanding overall. This early research indicates that an 'argument' based approach to explanation, that produces a more complex explanation for a response, appears to result in improved understanding.

This rationale and principle category coding scheme was later utilised by Berthold et al. (2009). The findings suggested that the rationale based explanations fostered both conceptual and procedural problem solving knowledge. This result may have occurred as constructing rationale style explanations involve deeper reasoning and interaction with the material. This research examined the explanations broadly in terms of generalised

statements but not the explicit structure and coherence per se and how these might relate to the positive outcomes.

Generic approaches to analysing argument quality are often developed specifically for research objectives and particular domains. Sandoval and Millwood (2005) suggested that quality interpretation for arguments and explanation resides within the domain in which they are constructed. In other words, quality depends on the application of the data within them and the sophistication of the interpretation and use of the data. The study employed a content analysis approach to evaluate scientific arguments produced by students in terms of the appropriate use of scientific concepts and the quality of interpretation of the evidence. A scoring scale was developed to label the level of 'scientific' data proposed as support for a claim and a framework of five types of rhetoric were identified.

The simplest level of rhetoric, labelled 'inclusion,' referred to explanations that merely included data without any further analysis. The second level -'Pointer' - directed the reader (e.g. "see diagram") to where data was situated without describing its relevance to the claim. The third level labelled 'description' referred to explanations that described data in terms of what it "said" without linking the data fully to the argument. The fourth level – "assertion" – claimed that the data "showed" evidence for the claim without fully explaining how. The final and most sophisticated level of rhetoric was "interpretation" which referred to explanations that pointed out specific aspects of the data and how these offered support for each claim. This type of quality argument analysis predominantly assesses knowledge quality and manipulation of scientific data specifically.

3.2.2.4 Use of Elaborations

As discussed above, certain types of explanation and lengthier explanations have been shown to be more effective in facilitating learning. Explanations and arguments have also been deconstructed into smaller components that may indicate quality by way of enhancing learning. One particular type of component that has been examined in research is that of elaboration. The use of elaborations has again been linked to learning performance.

Research by Schworm and Renkl (2007) examined the positive correlation between the number of 'elaborations' given in response to a physics based problem solving task and learning outcomes. Elaborations were the number of individual and distinct statements presented in an explanation and were coded into one of five categories in the study. The categories were not organised in order of complexity or quality and so were considered

equal in this sense. Explanations were segmented and an elaboration label applied to each segment. The first category of elaboration involved making connections between concepts presented in the material. The second, referred to structural aspects of the learning material.

The third category of elaboration referred to an instance of mathematical content relevant to the question. The fourth elaboration category was assigned if the segment made out of context comments and the final category, labelled metacognition, referred to personal opinion regarding approach to the task. The total number of the categories found in the explanation were taken to be the number of elaborations contained within the explanation. It was suggested that the quality of an explanation could be determined by the number of elaborations given. Thus the number of elaborations within a self-explanation was a good indicator of the learning processes undertaken. However, this is a not a great deal more insightfull method of analysis than purely examining explanation length, without a thorough differentiaton of the types of elaborations used and how these may impact learning.

The production of elaborations on learning materials is also known as inferencing (Chi, 2000) and it facilitates declarative knowledge and consolidates links between concepts. 'Elaborations' were defined by Chi (2000) as explanations on why a particular concept had been chosen (debatebly, this could also be considered a 'justificatory' argument). A further interesting finding in this work is the lack of positive learning outcomes observed for vocalised explanations. This again suggests that there is something unique about a written explanation compared to a spoken explanation and that explanation and argument may be indicative of cognitive processes. The research suggests that vocal explanations are more akin to working memory dumps and not as concerned with coherence in comparison to a written response.

Research carried out by Wolfe & Goldman (2005) also identified elaborations as key components within explanations. An investigation was conducted to examine how students process alternative texts using think aloud prompts. Paraphrases and elaborations were the most common types of activity. Elaboration involved connecting the information to prior knowledge as well as information both within and across texts. The complexity of reasoning in response to questioning was predicted by the think aloud comments that increased the coherence of the texts. Those comments that utilised knowledge and produced connections throughout the surface text were the most beneficial. This indicates a processing strategy

that is used to make sense of conflicting accounts. In contrast, paraphrasing does not generate connections to prior knowledge or allow for interpretation.

Rottman and Keil (2011) also suggested that elaboration provides a functional role pertaining to argument strength and persuasive attributes. Elaboration may be a cue to importance. If an element had been elaborated upon it was perceived as important enough to focus time and effort in explaining further. Elaborations could also be considered a more complex type of explanation in certain contexts. The production of an elaboration may be an indicator of prior knowledge and understanding that goes beyond a paraphrase. Elaborations may help to fill knowledge gaps or extrapolate concepts. As discussed previously in section 2.3, persuasion is another dimension on which argument quality can be assessed in terms of the strength of the argument impact on a receiver's position.

Overall, due to the varied nature of the methods within the research discussed above and the lack of concordance in the analysis methods, these research specific approaches, that ascertain quality based on features which correlate with desired outcomes, do not provide robust clarification of which particular structures within an argument impact upon the decision making process and how these features can help or hinder decision confidence and performance in such tasks.

3.2.3 Domain Independent Structural Quality Analysis

3.2.3.1 Introduction

In contrast to some of the research specific methods described above, this section will describe some of the most popular structural approaches argument quality analysis, for which quality is not considered related to outcomes. These methods are grounded in theoretical knowledge and are applicable to a wide variety of domains. A summary of the approaches discussed can be seen in Table 3. This discussion will provide a foundation for further investigation and the development of improved argument models and analysis methods.

The topic of argumentation as evidenced in the discussion so far, is inherently complex as it has applications in highly varied aspects of human behaviour, from legislation to medicine, and education to behaviour change. The complexity of context and approach to the study of argument has resulted in somewhat disconnected research and inaccessible methodology in some respects.

A unified theory of argument that is applicable across domains and offers a rich sense of argument context and purpose is currently elusive in the HCI community or indeed within social psychology; however there are a variety of well referenced and informative attempts to construct a domain general model of argument, some of these have been discussed in the previous section. The use of models enables judgements to be made about argument quality in less established domains such as the social sciences and more controversial areas where there may be disagreement as to what normatively 'correct' ideas are. Therefore, an argument could be judged on its composition from an analytical perceptive, rather than its inherent 'correctness.' In well-structured domains such as thermodynamics, the strength of an argument and its normative aspects may be much more tangible and measurable. In the case of less structured domains, argument models may be used to supplement other approaches such as a knowledge assessment in order to provide a more informative and holistic view.

Authors	**Quality Indictor**
Reed and Walton (2007)	Critical questions for argument schemes – test validity of argument and speech event.
Kuhn (1991)	Dialogue based theory – quality determined by facilitation of argument and presence of rebuttals.
Buckingham Shum and Okada (2008)	Modal qualifiers (Toulmin) as indicators of argument strength e.g. use of 'mostly'
Von Aufschnaiter, Erduran, Osborne & Simon (2008)	Use of warrants (Toulmin) to validate an argument.
Osborne et al (2004a)	Toulmin based hierarchical quality scheme with extended use of rebuttals as indicator of quality.
MacLean, Young, Bellotti, & Moran (1991)	Design Space Analysis – quality determined by presence of Questions, Options and criteria considered.
Petty & Cacioppo (1984)	Persuasive strength – measured by pretesting arguments (no explicit structural indicators).

Table 3 Summary table of structural argument quality indicators

3.2.3.2 Walton's Argument Schemes

A popular framework for argument modelling and quality analysis are the 'Argument schemes' developed by Reed and Walton (2007). The argument schemes comprise of a comprehensive list of argumentative types by which arguments can be analysed and compared. This dialogue centred theory asserts that each type of dialogue can be assessed as an argument that has an intended goal in terms of the impact on the receiver and the facilitation of argumentative discussion.

The schemes enable arguments to be categorised and then evaluated by an analyst based on the compliance with a series of critical questions relevant to the chosen scheme. The schemes are simple representations of types of arguments that people make, along with critical questions to ascertain the strength of the argument presented. One such scheme is that of 'argument from example,' (see Figure 6) whereby a claim is supported with an example in which it holds true. Many students use this type of argument and often use personal anecdotes in place of the example to support a claim (Nussbaum, 2011). The use of a personal anecdote may then be considered as weak when applying the critical questions, as the evidence is not generalizable. The critical questions do rely on the subjective interpretation of the person evaluating the argument, but offer some reliable indictors of quality that could be used in assessment (Nussbaum, 2011).

"Argument from example: An example is used to support a generalization."
***Critical Questions**:*

1. Is the proposition presented by the example in fact true?
2. Does the example support the general claim it is supposed to be an instance of?
3. Is the example typical?
4. How strong is the generalization?
5. Are there special circumstances in the example that impair its generalizability?

Figure 6 'Argument From Example' Scheme

The argument schemes are considered in the context of dialogue. In this respect arguments are thought to act as a specific speech event. The six types of speech event that can be used to categorise and evaluate an argument by an analyst are presented in Table 4.

Label	Description
Persuasion	Resolution of a conflict of opinion to resolve or clarify an issue
Negotiation	To make a deal that is satisfactory to all participants
Inquiry	To find or verify evidence in order to evaluate a hypothesis
Deliberation	To decide the best course of action in a practical situation / choice
Information seeking	To acquire or give information
Eristic	To fight and quarrel without any reasonable goal

Table 4 Speech Event Types for Argument (Walton, 2000).

The extent to which an argument fulfils its purpose in terms of the speech event can be used as an indicator of quality in terms of its effectiveness within the argument dialogue. This type of quality is more difficult to determine is examining an argument constructing by an individual that has no immediate receiver by which the impact can be measured.

Overall, the list of argument schemes is particularly extensive and rigid; therefore it is sometimes difficult to effectively address the critical questions for each scheme in addition to assessing the speech events, particularly if in depth knowledge of the context or subject area is not readily available to the analyst. This theory is more centred on examining argument in the context of a dialogue and may be less informative to apply to a standalone, argumentative piece of text, which is a common type of activity within formal assessments in education.

In spite of its criticism, Walton's dialogue theory is considered useful for providing a range of specific criteria that could be incorporated into scoring schemes or for use in qualitative data analysis of how students organise responses and determine which augment schemes are being used. Similarly, many arguments may be comprised of numerous argument schemes, and this may then be an indicator of complexity in itself. This model may not lend itself specifically to the modelling of typical arguments, as the schemes do not have intended domains attached to them, or an overarching coherence. A successful argument will only be defined as such if the critical questions are met, a process which will require subjective, human interpretation. No particular arrangement or set of argument schemes has been proposed as a model of typical or desirable argumentative structure in any particular domain.

3.2.3.3 Kuhn's Argument Quality Evaluation Framework

A similar coarse grained approach for assessing argument quality, again in the context of dialogue, was developed by Kuhn (1991). Kuhn's original investigation into argument analysis examined to what extent people are able to separate theory from supporting evidence or their own theories when producing arguments. The original assessment of argument quality proposed by Kuhn, Shaw and Felton (1997) suggested that arguments could be considered within a framework of three ('A', 'B' and 'C') categories. Category A arguments effectively consider alternatives to the main claim and category B and C arguments only propose evidence for one side of an argument with category C contained the weakest evidence. Both sides of the argument (pro and con) were also differentiated into 3 types; functional arguments (claims with evidence), non-functional arguments (claims without backing) and non-justificatory arguments (based on sentiment, or common belief).

Kuhn and Udell (2003) investigated how this dialogue based framework could be used to enhance and assess argument quality. Although the focus was predominantly on discourse, the framework was applied to arguments that put forth at the outset an 'argument' task prior to any interaction or feedback. Three aspects of argument skill were assessed in the dialogue. Firstly, the quantity of different reasons generated as part of an argument. Secondly, the quality of argument produced by the individual (based on the assessment scheme proposed by Kuhn, et al. (1997), and finally the quality of argumentative discourse a participant produces in dialogue. Each of these dimensions was assessed at the outset of the study and again following an interactive discussion. The quality of argument in discourse could be analysed by applying one or more of 24 specific labels to code large segments of the dialogue (Felton & Kuhn, 2001). The full list of labels is available in Appendix 1. These codes were applied to larger segments of discourse and labelled with a category depending on their function such as 'agree,' 'question' or 'counter'.

From this combined research Kuhn suggested that four components were essential to any successful argument model. The four parts comprise of a statement of theory and evidence, a statement of alternative theory, a rebuttal of the alternative theory and a final counterargument and rebuttal. The work suggested that some strategies for argumentation were more powerful than others in the context of dialogue, such as 'counterargument.' Powerful in this context refers to the persuasive nature of the argument and the increase in argumentative skill required to effectively deliver this type of strategy. However, persuasion

is fundamentally determined by the impact on the receiver which may be influenced by the current position and openness to change, rather than explicit structural properties of the arguments.

The Kuhn dialogue based framework has been used in to determine argument quality in various collaborative contexts e.g. Hollingshead and McGrath (1995) and Marttunen (1998). Joiner, Jones and Doherty (2008) assessed the quality of student arguments produced in an asynchronous collaborative setting Quality was assessed initially in terms of how effective the argumentative elements were in facilitating speech and dialogue. This method was supplemented in a second investigation which also examined the argument content using a Toulmin based quality scheme, developed by Osborne, et al. (2004a). This dual approach helped to enrich the analysis by providing more detailed insight into argument content along with Kuhn's more dialogue based approach. This quality scheme will be discussed in more detail in section 3.2.3.4.

This framework has strong similarities to the Toulmin model of argument which is somewhat easier to utilise due to its intuitive wording. Both approaches involve a similar manner for identifying broad structures of argument elements through subjective judgement, though there are evident important differences in the context of use. Kuhn's frameworks are dialogue focussed and evaluates argumentative elements based on their function in relation to other elements and how they help to progress an argument. Therefore, these frameworks may not be appropriate for analysing the typical standalone arguments that are generated in educational and particularly, individual e-learning environments. The empirical work which underpins the framework is based upon face to face communication. This perhaps hampers the use of the framework in the emerging fields of online or computer mediated work. The Toulmin model, which is discussed in more detail in the following section, is not dialogue dependent (although it is considered dialectical) and omits any attempt to suggest formal, interaction based properties to argument function.

3.2.3.4 Toulmin Based Quality Analysis Approaches

The Toulmin framework outlined in section 3.1.2 has been incorporated into numerous approaches that aim to assess argument quality. This is a strategy adopted by Buckingham Shum and Okada (2008) in the Scholarly Ontology framework which focusses on the use of modal qualifiers within an argument as an indicator of strength. The framework was designed to determine argument quality by assessing the strength of the claims made within

the argument. Broad argumentative elements are labelled and weights are applied to these based on the positive or negative terminology within them. This is problematic as the weights offered by analysis models (that typically represent polarity) include suggestions that use of the word 'supports' may be stronger than 'agrees.' Similarly, 'unlikely to affect' would be considered weaker than 'prevents.' Consequently the Scholarly Ontology would suggest that a weaker modal qualifier would be not as strong as a statement of absolute certainty, however in terms of persuading the reader, the 'weaker' qualifier may in fact be more desirable as it may mean the argument falls into the latitude of acceptance if the position is not completely polarised. Essentially, although modal qualifiers can signal weaknesses, they may also soften an argument to make it more palatable to the reader. Additionally, the use of wording in these respects is also subjective on the side of the author and may have no bearing on the actual quality of evidence used. This presents a problem when using these types of qualifier elements as indicators of power or 'weight' without independent assessment of an analyst to determine how appropriate they are.

Most attempts to use the Toulmin model to assess argument quality appear to focus on the use of supported warrants. For example Von Aufschnaiter, Erduran, Osborne and Simon (2008) used the Toulmin approach to analyse the verbal conversations of school pupils to assess their use of backing and warrants. This is fundamentally problematic however, as most warrants are often not directly signalled in the arguments. This classification approach to using the Toulmin model does enable fairly systematic deconstruction of arguments, but the assumptions of argument quality or skill must be made by the analyst and determined by the argument context and purpose. For the purposes of ascertaining the persuasiveness of an argument, usually the Toulmin model (or indeed any analytical model) would need to be supplemented by an assessment of evidence relevance and accuracy.

The Toulmin model framework is often used in conjunction with other quality analyses to further support any conclusions drawn from its use. An example of this approach can be found in research by Ferretti, MacArthur and Dowdy (2000). This research utilised the Toulmin model elements to analyse the structure of arguments, presenting the frequencies of each element as an indicator of complexity. In combination with the Toulmin based analysis a 'persuasiveness' scoring rubric was developed. The rubric rated arguments on a 0-7 scale, with '0' being a response with no personal opinion and '7' being an elaborated argument that is well organised and concludes effectively. The rubric required that holistic judgements were made by an analyst about the overall persuasiveness of an argument. These two measures combined were considered to give a strong indicator of 'quality.'

However, these measures lack indicators of ordering or a guide for the integration of elements, only asserting that the element existed. There is little sense of mapping between the Toulmin model elements and the persuasiveness rubric scoring. This is perhaps a result of the Toulmin model lacking in a suggestion of power for the elements and little indication in the original proposals of which elements and structures form the more powerful arguments.

The Toulmin model has been criticized for being difficult to use as an evaluative tool on this basis (Duschl, 2008). The model was developed to describe general informal arguments but provides little indication of how these may be organised depending on certain contexts, which elements may be particularly concerned with persuasion or confidence in a decision, or which are more pertinent to perceived argument quality. In its defence, it does have a field independent nature and can be applied with relative success to any domain.

In order to begin to address the requirement for a more robust and defensible Toulmin based quality scheme, some researchers have attempted to assign levels to the argument components. One particular group of researchers, Osborne et al (2004a), developed a framework based on the Toulmin model to assess the quality of argumentation present in science arguments. The quality aspect was centred on argumentative structures and independent of an assessment of the quality of data used. Each statement within the argument produced by each individual would be assessed by level and type of argumentation (claim, warrant, backing, and rebuttal). This framework was referred to as a 'quality scheme' and represented a systematic way to label the strength and progressive complexity of arguments.

	Description
Level 1	Consists of arguments that are a simple claim versus a counterclaim or a claim versus claim.
Level 2	Has arguments consisting of claims with data, warrants, or backing, but do not contain any rebuttals.
Level 3	Has arguments with a series of claims or counterclaims with either data, warrants, or backing with the occasional weak rebuttal.
Level 4	Shows arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counterclaims as well, but this is not necessary.
Level 5	Displays an extended argument with more than one rebuttal.

Table 5 Argument Quality Analysis Framework taken from (Osborne, et al., 2004a).

The framework was intended to provide a method for distinguishing between arguments based on argumentative features considered critical for good arguments in a science context. It is fundamental to the discipline that scientific ideas are required to be supported by evidence, which can be strengthened by the use of qualifiers, rebuttals and alternatives. The authors wished to focus on the argumentation used as a foremost indicator of quality and evidence of argumentative skill, as opposed to a classic approach of examining knowledge quality. This should uncover the critical argument processes that occur as an indicator of argument competency and skill, rather than the ability to reproduce accurate evidence. This framework utilised the Toulmin elements and proposed a hierarchy of argument quality (Table 5) based on the frequency of specific components.

Examples of the types of arguments that could be assigned each quality level are shown in Table 6. The simplest level of argument consists of a simple claim with no supporting evidence or backing. The second level has a claim with some valid data and backing. The third level of argument introduces the possibility of a weak rebuttal, an attempt to invalidate or acknowledge an opposing claim. The fourth level of arguments should contain backing and clear rebuttal with supporting evidence. The fifth and highest level of argumentation should contain at least two clear rebuttals with backing to represent a wider consideration of the argument space.

Quality Level	Example Argument
1	I agree with the argument of pro-GMF, because it is more persuasive.
2	A nuclear power plant needs large quantity of water for the cooling system. The discharge of warm water into the sea would cause a detrimental effect on the ocean ecology of the surrounding area. **(warrant)** So I oppose the construction of any nuclear power plant. **(claim)**
3	The reasons why I do not support GMF **(claim)** are: 1. The potential risk of GMF is not fully understood **(backing).** The argument of resolving food shortage by planting GMF is not realistic at all. In fact, food shortages in some areas around the world are mainly an issue of distribution. It can be solved through United Nations' operation. **(weak rebuttal)**
4	1. By 2008, the total area of growing GM plants is more than 125 million hectares. **(backing)** The GM golden rice containing much more beta-carotene than typical rice can be used in areas where there is a shortage of dietary vitamin A. **(data)** I do not agree with the statement- "Growing GMF is beneficial only for rich businessmen and not good for farmers". For example, currently there are 13.3 million farmers in 25 countries growing GM corns for their living. In other words, 90% of farmers in developing countries are dependent on GM plants. **(rebuttal)**
5	One of my reasons for not supporting GMF **(claim)** is that many GM plants were derived from their natural enemies' which may affect human bodies' tolerance to antibiotics. **(backing)** I disagree with the statement as it was found that one allergic compound was also transmitted into the soybean. **(rebuttal)**I would also challenge the statement-"Those who are against GMF always give the excuse that it has potential risks. However, no solid evidence was provided." In fact, at the beginning stage of any science achievement, scientists were not fully aware its potential detrimental effects. The impacts on aquatic life of DDT and bird population were not known in the 1950. **(rebuttal).**

Table 6 Examples of arguments at each quality level adapted from Lin, et al (2012a).

Most notably (and congruent with Kuhn's assertions on the power of counter argument), full rebuttals are the most salient feature of the highest quality arguments. Kuhn (1991) suggested the use of the rebuttals is ''the most complex skill'' in argument as it relies upon the demanding, higher level process of integrating ''an original and alternative theory, arguing that the original theory is more correct'' (p. 145). Therefore, the authors posit that arguments with rebuttals are of better quality than those without as counter claims without a rebuttal can become circular arguments with no resolution or change in attitude. Rebuttals were also considered to be indicative of a higher quality argument in terms of learning and skill acquisition as the process of constructing rebuttals requires evaluation of the validity and strength of available data. This is reflected in the hierarchy of levels in the framework.

The quality framework outlined above has been utilised by researchers seeking to evaluate argument quality in educational domains and to support argument skill acquisition. Lin, Hong & Lawrenz (2012) adopted the framework to analyse group constructed arguments in an asynchronous online communication context. The framework enabled comparisons to be made between collaborative contexts using the level system and was effective in identifying a measurable difference in argument quality between groups. Similarly, Yeh and She (2010) used a modified version of the quality framework, further deconstructing each component of argument into levels demonstrating argument quality improvement as a result of task intervention.

The framework has also been validated as a tool for improving argumentative skill (Okumus & Unal, 2012), teaching science argumentation (Simon, Erduran, & Osborne, 2006) and assessing argument in other contexts (Simon & Johnson, 2008). It enables the quality of argument to be comparably assessed by identifying the number of components and thus the level of complexity. The framework has since been used in science education research as a tool for both assessing and supporting quality in argumentation (Yeh & She, 2010).

An advantage of this approach to quality assessment over functional types of argument quality evaluation i.e. Kuhn's evaluation framework, is that the structure of the argument itself and its components are the focus rather than its function in the context of dialogue, external persuasion or perception of validity. However, the framework focusses on a single dimension of complexity and quality, that of the use of rebuttals. There are indeed a number of other dimensions along which argument can be evaluated including the receiver's position and the quality or validity of the knowledge within.

In addition there are a number of practical limitations in using the framework including the ambiguity in identifying the components. Some aspects of argument are often implicit, particularly if the material used is not accessible by the receiver. Additionally, the focus on argument structure disregards the validity of evidence used.

Most research that has used the Toulmin model to construct argument quality frameworks is still fragmented in its approach, with various researchers developing different strategies. Generally, little has been agreed upon as to what constitutes a 'strong' versus 'weak' argument as this is still considered to be largely domain dependent. There is still an evident need for agreed frameworks that focus on argument construction as the primary indicator of skill and argument quality, with an assessment of the evidence used as a secondary aspect.

3.2.3.5 Design Space Analysis Framework

A well-established strand of research into argumentation that moves away from the traditionally dialectical function of argument has arisen from HCI design research. In this domain the importance of justificatory arguments and the ability to construct a balanced rationale is considered important in the context of design justification. There is a need to encourage the effective production of rationales for a design solution and analyse existing arguments effectively. The rationales generated in the context of software design can assist in redesign of products or in reuse (Burge & Brown, 2003). Although these arguments are not necessarily intended to be persuasive, they need to be comprehensive and demonstrate that possible alternatives have been considered, for the purpose of completeness and for possible future use.

One approach proposed to structure rationales in the context of design is that of 'Design Space Analysis' (DSA). This approach can act as a framework during development or as a post hoc structuring tool using the design documentation (MacLean, et al., 1993). The framework (summarised in Figure 7) suggests that an effective rationale should consist of three components; Questions, Options and Criteria (QOC). The QOC (or DSA) framework is a useful notation format to construct or extract a rationale from design documentation (MacLean, Young, Bellotti, & Moran, 1991). This type of rationale is a different type of argument to those usually examined in research.

A DSA approach examines rationales as a standalone type of argument, in contrast to the usual discourse based argument frameworks. The primary function is to justify rather than persuade and to clarify how the best solution for a design was reached in the face of alternative options. DSA can help to structure a rationale for a design from the start of the design process using the three core elements. The 'questions' aspect identifies key design issues while 'options' indicate possible answers to the questions and finally the 'criteria' aspect states the features or constraints for assessing and comparing the options.

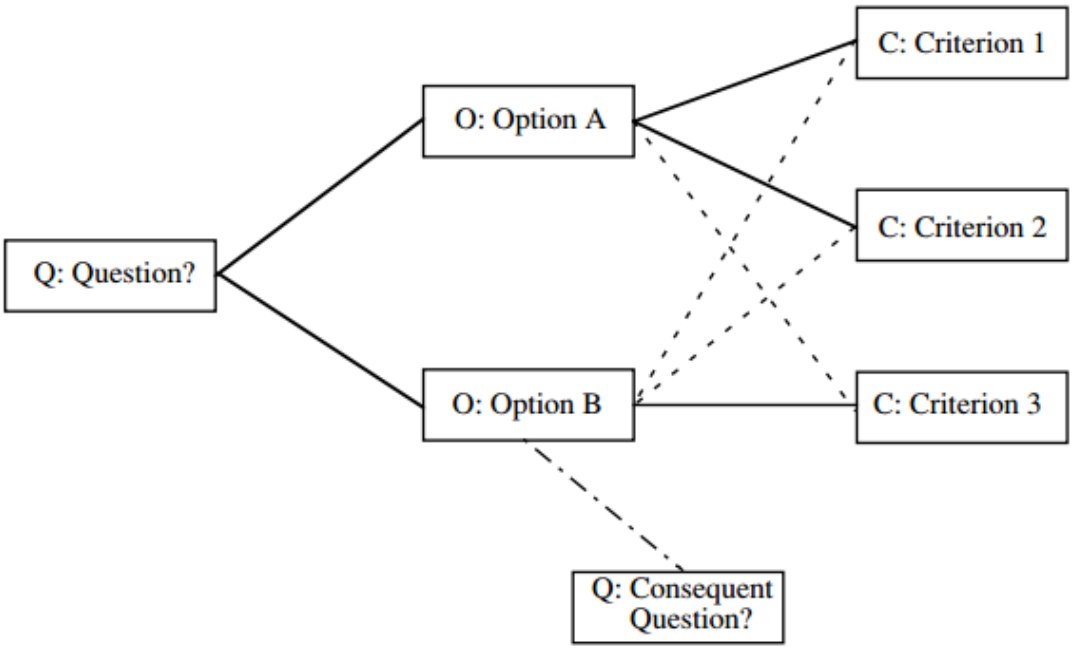

Figure 7 Design Space Analysis framework example taken from MacLean, Bellotti, & Shum (1993)

The use of DSA is important for sense making of a design solution and the reasoning behind it. It provides a bridge consisting of argumentative reasoning between the problem scenario and the proposed solution. There is a growing body of research that has utilised the QOC framework to construct a design 'ontology' which can aid machine mining and reconstruction of design rationales from existing unstructured documentation and reconstruct them in the QOC format. The aim of these systems is to minimise the cognitive load and additional work on the part of the designer and perhaps eliminate the need for explicitly constructing a rationale as part of the design process (Zhang, Luo, Li, & Buis, 2013). It is not clear whether this framework is applicable in other domains to represent a complete or well-structured argument, but if systems have been developed that can analyse free text

and reconstruct arguments based on this format, it may hold some applications in areas of natural language processing.

3.2.3.6 Pretesting Argument Strength

As an alternative to the methods discussed above, one particular approach that may offer a generic solution to argument analysis that is not grounded in any particular theory, domain or structural approach is that of pre-testing. Green (2007) suggested that pretesting arguments for their strength, by getting participants to rate how convincing they were, was an effective strategy of assessment argument quality. The Social psychological literature on persuasion (Petty & Cacioppo, 1984) would appear to support this approach. However pretesting does not offer specific insight into exactly why certain arguments are considered convincing by a reader and whether this may rest in part, on the structural features. It is also problematic as it centres on persuasion as being the main measure of argument quality as a function of its ability to convince a reader. As discussed in section 2.3.2, this persuasive power can vary depending on receiver involvement in the argument and where the claim falls in the latitude of acceptance. The main benefit of the pre-testing method is that is takes argument analysis out of the hands of the researchers, which does save time and resources.

3.2.3.7 Summary

The research specific approaches to assessing argument quality discussed above often depend on the goal of the activity in which they are situated. For example, if the goal is to facilitate conceptual learning, a good quality argument may be one that incorporates a significantly higher proportion of backing information or elaborations (Chi, 2000) to increase the amount of information processed, thus aid recall. However, some research suggested that a more reason based approach that includes a discussion of why a particular decision was undertaken could also enhance conceptual understanding (Roy & Chi, 2005).

In contrast, the more domain general approaches to argument quality analysis, such as the Toulmin based quality scheme (Osborne et al, 2004) and Kuhn's dialogue based framework, focus on the use of rebuttals as a measure of the breadth of the argument and perhaps a wider consideration of the alternatives to a viewpoint. Similarly, Design Space Analysis requires that the available options needed to be considered in a design rationale in order for it to be most useful to others in the future. In this respect, a rationale needs to consider the alternatives to the claims made in order to be useful in a collaborative context, perhaps to

fully communicate a well thought out decision, and perhaps to persuade others to concur. Additionally, these approaches suggest that good quality arguments are those that adopt a two sided approach, and pre-empt any opposition to the claims made. These methods are arguably easier to apply to a wide variety of domains, particularly in exploratory investigations.

The thesis investigates the aspect of intended direction and how this perception may influence the quality of arguments produced. The research suggests that argumentative aspects such as rebuttals may be more common in other directed arguments (Lin & Lehman, 1999). On this basis, the quality analysis approach in this thesis for will incorporate the methods that include rebuttals and considerations of alternatives as the quality aspects and propose that rationales that adopt a two sided approach are of greater quality than those that do not. Additionally, the research that examines aspects of quality that relate to learning performance indicates that a more justificatory approach, that could include the use of rebuttals, may be beneficial. The next chapter will discuss the currently available systems, many of which utilise the frameworks discussed here, which aim to support and evaluate argument construction.

4 Technological Argument Support and Analysis

4.1 Introduction

The previous chapter explored the complex and varied approaches to argument modelling and analysis and some of the contextual constraints that are worthy of consideration.
In recent years a number of these argument models have given rise to systems that support argument construction and evaluation. Section 4.2 will describe some of the systems that support individual and collaborative argument construction and some of the steps that have been taken towards computer supported argument analysis will be outlined in section 4.3. Following this, the apparent challenges that argument support and analysis pose for systems design will be discussed in section 4.4. Finally, the concluding statements for the literature review will be outlined in section 4.5.

4.2 Types of Argument Support Systems

Technology and education in particular are fast becoming inseparable partners. The role of technology in facilitating and motivating argument and argumentative skill acquisition in students has inspired a wealth of research. Argument itself is a central theme in many learning domains, used both to enhance understanding of conceptual knowledge by increasing evaluative and critical approaches and to increase the skills of effective argument in themselves.

Argument support systems are frequently founded on argument ontologies, such as those by Toulmin and Walton discussed previously. There have been a number of attempts to produce systems based on these frameworks that both support and teach argumentation as a skill e.g. see review paper by Scheuer, Loll, Pinkwart and McLaren (2010). These systems are designed to support and analyse arguments as they are constructed. These systems are borne out of need to enhance a skill that although used daily, is often lacking as many people struggle to produce effective arguments or confidently evaluate the arguments of others. This is particularly pertinent in ill-structured domains such as the social sciences, where producing rationales for viewpoints in the face of contentious evidence is at the core of the discipline. In the past decade, at least four different design approaches have been adopted to develop tools to support and analyse argumentation (Scheuer, et al., 2010). These approaches are summarised below:

1. *Scaffolding arguments;* with prompts such as sentence starters. Have been shown to have positive impact on the quality of arguments produced (Schwarz & Glassner, 2007).

2. *Graphical representations of argument;* have been shown to improve collaboration (Nussbaum, Winsor, Aqui, & Poliquin, 2007) and problem solving (Easterday, Aleven, & Scheines, 2007).

3. *Assignment of argumentative roles in collaborative argumentation;* such as 'moderator' 'summarizer' and 'evidence searchers' (Schellens, Van Keer, De Wever, & Valcke, 2007).

4. *Argument quality support approaches;* provide feedback on actions and solutions created. Hints and recommendations guide students during the task (Pinkwart, Aleven, Ashley, & Lynch, 2006).

Graphical representation approaches have become the most prevalent type of design developed and supported by many educational research groups. The systems developed are referred to as 'Computer Assisted Argument Mapping' (CAAM) tools. Popular examples of CAAM Tools include Compendium (Selvin et al., 2001) Acuriaca (Rowe, Macagno, Reed, & Walton, 2006) and Rationale (Van Gelder, 2002). These graphical representation systems aim to support and improve individual argumentative skill as a direct result of constructing an argument using the system. These tools have predefined elements that comprise an argument that can be selected and completed by students to produce a visual map. These argument maps clearly demonstrate the pros and cons, highlighting specific evidence supporting each side. Examples of developed argument support systems that utilise the various approaches along with their ontological basis are summarised in Table 7 and Table 8.

The review of the argument systems literature conducted by Scheuer et al. (2010) identified a number of further distinct types of argument construction types that are most prevalent within the available systems. Argument systems can either adopt modelling or discussion oriented strategies. Systems that focus on modelling are typically rigid in both input methods and output observed with a predefined, prescriptive structuring. On the other side, discussion systems utilise far less coherent text, and as a result are much less rigid, with sentence openers and light restrictions on text. User interaction varies widely within the

argument support tools discussed, with users constructing many different types of scaffolded and unstructured arguments. Systems that utilise free form arguments, that do not restrict argument layout, are the most useful in developing individual argument generation skills and self-reflection, as resources need to be sought independently to support arguments (Scheuer, et al., 2010). This type of argument is the most common format within the self-explanation research (Chi, 2000). Some systems can encompass both types and therefore be used to support individual argument construction and collaborative argument (Schwarz and Glassner, 2007) but arguably, most systems could be used with a shared screen.

The earliest argument support tools utilised a text based approach to argument. While this form of representation is easiest to implement in collaborative systems, it does not explicitly demonstrate the coherence or flow of an argument, although it is debatably a more intuitive approach for users. Most tools use a diagram based representation of arguments to guide the structuring of arguments as they are made. Some systems, Belvedere for example, are fairly unrestricted in the structuring of arguments, whereas some systems such as Araucaria are very rigid in their organisation, forcing users to adopt the prescribed argument structures.

Other types of argument construction within systems include those that prompt users to deconstruct and evaluate pre-existing arguments. This approach is demonstrated most successfully by the LARGO system (Pinkwart, Aleven, Ashley, & Lynch, 2007). Argument support systems are most often designed depending on the objectives of the task being supported. Objectives can vary from acquiring conceptual knowledge, to fostering argument skill development and supporting collaborative discussion.

Aside from the type of input the systems require, either structured or unstructured, argument support systems can additionally be subdivided into categories depending on their intention to support individual or collaborative argument.

Table 7 Summary of collaborative argument support and analysis systems with ontology and argument representation approach (adapted from Scheuer et al., 2010).

Collaborative Argument Support Systems					
Tool	***Reference***	***Ontology & Representation***	***Tool***	***Reference***	***Ontology & Representation***
AcademicTalk	Mcalister, Ravenscroft, & Scanlon (2004)	Dialog game, sentence openers	Aquanet	Marshall, Halasz, Rogers, & Janssen Jr (1991)	Hypertext labelling of components - Toulmin
Argument Web	Bex, Lawrence, Snaith & Reed (2013)	Mapping, AfD	ArguNet	Schneider, Voigt, & Betz (2007)	Mapping of supportive and opposing arguments
Belvedere v1- v4	Suthers et al (2001)	Expert knowledge model comparison, map	BetterBlether	Robertson, Good, & Pain (1998)	Sentence openers
CoChemEx	Tsovaltzi et al (2010)	Collaboration oriented scripts for discussion	CoFFEE	Belgiorno et al (2008)	Graphical and threaded discussion
Collaboratorium	Klein & Iandoli (2008)	Threaded discussion - argument map	Collect-UML	Baghaei, Mitrovic, & Irwin (2007)	Problem, answer solution to object oriented design
CoPe_it!	Karacapilidis & Tzagarakis (2009)	Knowledge map and argument labels	CycleTalk	Kumar, Rosé, Wang, Joshi, & Robinson (2007)	Scaffolded text discussions
DebateGraph	www.debategraph.org	Knowledge visualisation	Debatepedia	Wells, Gourlay, & Reed (2009)	Hypertext representation of pro-con text arguments
DREW	Corbel et al (2003)	Argument diagrams, chat	Epsilon	Goodman et al (2005)	Sentence openers – chat
Epsilon (interaction)	Soller (2004)	Sentence openers – chat	Group Tutor	Israel & Aiken (2007)	Sentence openers
Hermes	Karacapilidis & Papadias (2001)	Structural text prompts	IBIS/gIBIS	Conklin & Begeman (1988)	Chat, sentence openers
Interloc	Ravenscroft & McAlister (2008)	Sentence openers	LASAD	Loll, Pinkwart, Scheuer, & McLaren (2012)	Diagrams, highlight patterns
Questmap	Carr (2003)	Hypertext labelling of legal arguments	Room 5	Loui et al (1997)	Graphical prompts for argument progress
TC3	Munneke, van Amelsvoort, & Andriessen (2003)	Argument mapping			

Table 8 Summary of individual argument support and analysis systems with ontology and argument representation approach (adapted from Scheuer et al., 2010).

Individual Argument Support Systems					
Tool	***Reference***	***Ontology & Representation***	***Tool***	***Reference***	***Ontology & Representation***
Araucaria	Reed & Rowe (2004)	Argument mapping, (Toulmin, Wigmore)	ArguMed	Verheij (2003)	Template based text argument structuring
ARGUNAUT	McLaren et al (2007)	Ethical discussion support	Athena	Rolf & Magnusson (2002)	Argument mapping
AVER	van den Braak & Vreeswijk (2006)	Argument mapping – evidential reasoning	AVERs	Bex et al (2007)	Argument mapping – evidential reasoning
Convince Me	Ranney & Schank (1998)	Argument mapping	Carneades	Gordon, Prakken, & Walton (2007)	Expert models to evaluate legal dialogues
Compendium	Buckingham Shum & Okada (2008)	Argument mapping (Toulmin, Walton)	Digalo	Schwarz & Glassner (2007)	Micro-scripting, graphical tool
iLogos	Easterday, et al (2007)	Argument mapping	LARGO	Pinkwart, et al (2007)	Expert knowledge model comparison
Legalese	Hair (1991)	Graphical hypertext tool for legal argument	Pedabot	Kim et al (2008)	Textual argument organiser – compares to past discussion
Rashi / Biology	Woolf et al (2005)	Expert knowledge model comparison	Inquiry Tutor Rationale	Van Gelder (2007)	Argument mapping
KERMIT	Weerasinghe & Mitrovi(2006)	Sentence openers	Reason!Able	Van Gelder (2002)	Argument mapping
SEAS	Lowrance, Garvey, & Strat (2008)	Evidential reasoning	SEURAT	Burge & Brown (2006)	Rationale structuring
Zeno	Gordon & Karacapilidis (1997)	Text labelling - Toulmin based			

As shown in Table 7 and Table 8, there is a fairly distinct separation between these intended support categories, although it could be argued that a screen could be viewed by multiple persons even if only one person at a time can interact with it. Individual systems are usually Intelligent Tutors which focus on supporting arguments that pertain to a particular domain, such as law or physics. However, collaborative systems usually evolve from Computer Supported Cooperative Work (CSCW) research, which tends to focus on the facilitation of good collaborative arguments as the task objective. It seems an interesting anomaly that such a large number of systems are designed with single user interaction in mind, when argument is an inherently interactive activity. This is an aspect which may need to be bridged depending on the goal of the successful argument, whether facilitating critical thinking and decision confidence, in which case a single user would be sufficient, or to improve persuasive argument skills, for which interaction with another may be necessary.

4.3 Automated Argument Evaluation

4.3.1 Introduction

The development of automated argument analysis and quality evaluation systems is still in its infancy. As such there are a limited number of systems that attempt to provide feedback on the quality of user arguments available to discuss at this point. This section will examine a number of these approaches and the corresponding levels of analysis and efficacy.

4.3.2 The GATE Framework

Wyner, Schneider, Atkinson and Bench-Capon (2012) developed a semi-automated approach to argument analysis in order to assess online product reviews in terms of argument plausibility and effectiveness. The work aimed to investigate justificatory arguments for item purchase, in terms of determining the possible power of the argument. The authors used the GATE framework (Cunningham, Maynard, Bontcheva, & Tablan, 2002) which would label premises (such as ‘because’) and highlight positive or negative markers (such as ‘increased’ ‘most’). Argument power is not necessarily determined by the length or number of claims, but rather the attempts it makes to address potential rebuttals and assess counter arguments. The tool also identified specific product features that were mentioned. The semi-automated approach developed is specific to the purchasing scenario that it was designed to evaluate. This is a characteristic of many evaluative systems produced in research that are consequently difficult to apply in other contexts.

4.3.3 The Belvedere System

One of the few systems to offer a way of evaluating the strength and quality of an argument as part of an argument support system is Belvedere (Suthers et al., 2001). Other systems could perhaps be argued to provide an implicit type of feedback by the presence of different argument elements within the tool which may imply to the user that a good argument would require, as a minimum, use of all of the available elements. However, it would not be clear how the nodes should be arranged or their inherent strength and whether more than one could be used. The Belvedere system provides a matrix table to address this, showing all of the argument elements and how they relate to one another. From this table users can see where there are gaps in their reasoning. The system will also provide on demand, pre formed messages based on syntactic elements identified in the text, to encourage students to think further about their argument. These prompts are not diagnostic of issues per se, as the aim is not for users to blindly implement advice, but to respond to and think about more general guidance on strengthening the argument.

4.3.4 The LARGO System

A more domain specific system for argument analysis is the Legal Argument Graph Observer (LARGO). The system was developed to assist with the construction of legal arguments (Pinkwart, et al., 2007). LARGO, like Belvedere, has also implemented an evaluative and feedback function to the system. The interface provides a graphical representation of arguments based on models of hypothetical reasoning. The system provides feedback on missing elements of the diagram or provides prompts for generalised improvement of the argument. The models are reminiscent of the Toulmin model argument structures, with nodes and relationships between claims and rebuttals denoted by connective lines. The driving aim of the system was to encourage better quality of argumentation than free text alone. In terms of argument support and learning gains, students using LARGO have yet to be shown to outperform students who use text based learning materials (Pinkwart, et al., 2007). This may be due in part to the naturalistic language representations being more easily understood and that forced graphical representation still represents an increased cognitive load above the argumentative activity itself. However, in spite of the possible limitations, LARGO has implemented the most comprehensive attempt to provide feedback on argument quality that has been achieved thus far in computer supported argument systems.

4.3.5 The G.A.I.L Argument Analyser

More recently Green (2013) demonstrated a prototype argument analyser system built within an existing learning environment. The basis for the analyser was a desire to provide feedback to students in response to their arguments, in a manner which enabled them to learn how to improve their argumentation skills. The system incorporated the Toulmin model elements as a determinant for the elements expected. The system checks whether the argument generated matches one of the acceptable arguments. These acceptable arguments, referred to as 'expert arguments' were initially modelled within the Genetics Argumentation Inquiry Learning (G.A.I.L) system. Weaknesses in the argument are then fed back to the user and are non-domain and non-problem specific. For example, if the warrant does not match the expert arguments the system will highlight an error for the user. The system also suggests critical questions, which are more generic but prompt deeper consideration of the argument. The appropriateness of the feedback has not yet been empirically studied or optimised for this system, only the potential to identify errors in the fundamental argument structure.

In summary, it seems in spite of the aforementioned attempts to automate argument analysis, there is still a considerable gap between natural language and accessible forms of automated argument evaluation. There are some generic automated approaches to text analysis available that reveal rhetorical features. Two of these approaches - strongly related to RST - will be discussed in the next section, with a view to presenting the parsers as potential candidates for argument structure analysis and evaluation.

4.4 Challenges for HCI research and Argument support

The previous sections have reviewed some of the developed systems for argument support and analysis. The current understanding of argument and design still has a number of challenges to overcome in order for a comprehensive approach for argument support and analysis to be developed. This next section will consider some of the most pressing issues for Human Computer Interaction research and design in the domain of argument.

A prohibitive issue in the development and research of argument support systems may be the lack of reusable source code. Very often, each new investigation into supporting arguments will construct a unique system. As a result of this, it is not clear from the research which tools offer the most effective methods to support argument construction, or which

visual representations are more appealing and useful for users in various domains. The evidence for this is that the apparent abundance of argument support and analysis systems listed in Table 1, is in sharp contrast to the number of systems freely available for use in research. In addition, there is yet to be an agreed overarching 'theory' of computer supported argumentation, which may offer a framework of guidance for design or system use (Scheuer, et al., 2010). Sharing source code would allow for a greater evaluation of systems and wider scope of studies outside of each individual research group. Some support systems are freely available to download and trial but the modification and adaption of inbuilt system features to fully investigate user experiences are not possible without access to the source code.

A further drawback of these argument systems discussed so far, is that they are more often, domain specific and the underlying analysis mechanisms and feedback provision is particular to the topic around which the system was designed to support. A possible solution to this is the introduction of a collaborative analysis approach to argument research that would build comprehensive theories and models of arguments via input from experts in a variety of disciplines. To encourage educators and students to embrace new argument support technologies there may be a need to create a simple web based interactive environment, that can be easily installed and accessed by all (Scheuer et al., 2010).

The Argument Web project (Bex, Lawrence, Snaith, & Reed, 2013) tentatively offers a solution to the disconnected approach of current argument research and includes new tools and systems that utilise the Argument Interchange Format (AIF) ontology for tagging and modelling arguments. The AIF ontology itself allows for arguments to be tagged with a machine readable marker (based on the Walton argument schemes) that could potentially be used for machine processing. The overall aim of the AIF is to begin to form a unified model of argument that can be truly domain independent. This interface allows any argument style explanations to be added to the database, with the aim of increasing expert analysis and the possibility of generating models. The Argument Web database offers access to a large corpus of arguments that is constantly being updated and analysed structurally by linguistic experts in various fields. The AIF systems require human analysis of the arguments, thus do not yet offer usable solutions to the issue of automated evaluation or feedback.

There are a considerable number of systems available that do claim to offer automated argument structure and quality analysis; however the most popular and widely researched of these have been discussed and it has been found that this functionality is still rather

generic and questionable in its efficiency. The issue of providing feedback and analysis of arguments for quality and coherence in most argument support systems is still largely left to the users, who may be able to assess an argument on the basis of completion level or rate how effective their argument may be. The systems tend to encourage users to think holistically about their arguments as opposed to primarily pin pointing structural and linguistic features. These systems employ human based feedback by implementing the use of strength indicators on the graphical nodes by which users can rate the strength of a claim or piece of evidence based on agreement or how convincing they think it is. If using the system collaboratively the contributions of peers can offer evaluative functions. Other systems have additional post argument analyses that require human moderators in order to subjectively assess novice arguments (Suthers, Weiner, Connelly, & Paolucci, 1995). Few tools aside from Belvedere, LARGO and GAIL have attempted to offer automated evaluation or feedback on argument quality.

The ontologies that structure argument support systems are often primarily based on the Toulmin or Walton argument schemes. The ways in which arguments are represented to users need to be intuitive. The representations need to be accessible and straightforward to manipulate to ensure that the cognitive load exerted by structuring an argument in the environment is not increased to a detrimental extent, by providing confusing or frustrating representations. One of the most logical ways to overcome this would be to prompt natural free form text and build up the argument representation and evaluation from there.

In contrast to the specifically developed argument support and feedback systems, the type of linguistic labelling approach that RST offers is applicable to natural language and provides a richer form of analysis. The most common form of argumentative dialogue online for example is free text, and arguments are rarely presented intuitively in graphical form. This is worth considering as arguments are usually constructed as passages of coherent text in educational settings. Therefore the tendency of argument support systems to focus on producing diagrammatical representations of argument is not necessarily the most accessible approach. The graphical representations may in fact produce further cognitive load beyond that required for constructing an argument and the diagram itself may become a distraction. Similarly, the presence of nodes in graphical argument representation may also be a hindrance as the separate nodes may result in a disjointed view of the overall argument. There is not always space or prompting to include connecting phrases between the nodes, aside from the relational arrows between them. In contrast, free text can encourage a more

holistic focus on the overall flow of the argumentation as the 'node gaps' that are seen in the graphical representations are not evident in a free text approach, thus are intuitively filled.

The focus of explanation research on structured domains may be a result of ill-structured domains being inherently more difficult to control or to design specific features for. For example, using a web based interface to assist in analysing learner arguments, (Jermann & Dillenbourg, 2003) investigated the optimal learning goals in a web based domain. In terms of designing for learning, the task could be relatively simple when the goals are procedural skills, e.g. learning by doing, however, difficulty arises when the learning objective is the acquisition of abstract conceptual knowledge such as laws and theories as this is not straightforward or informed by the objective itself. It is therefore considerably more difficult to implement these types of goals involved in argument, tangibly and explicitly into an interface.

The lack of accessible and modifiable systems mean that using research findings across domains may be problematic as it is clear that argument is influenced by many contextual factors that need to be considered and supported. This has resulted in automated argument quality analysis and feedback systems remaining a relatively novel area of research. Argument quality analysis frameworks need to be adaptable, accessible and expedient for researchers to adopt in their work in order to study naturally generated arguments. Free form text as opposed to prescribed drop down menus and graphical models may be the most intuitive way to elicit arguments, particularly from novice (less competent) arguers. This thesis will attempt to take a step towards assisting in this area.

4.5 Concluding Statements

Through the discussion of the research presented in the previous chapters, it becomes clear that there are a number of perspectives from which argument evaluation can be viewed, and the numerous benefits that good argumentation can produce, such as enhancing knowledge, attitudes and argumentative skill.

The scope of argument analysis can be broadly categorised into the assessment of quality and the analysis of structure. Quality based approaches evaluate the strength and validity of supporting data contained within an argument or how the extent to which the argument supports the activity in which it is situated, such as task information processing or recall. Structural approaches consider the components that may be present in the argument and have also been adapted in quality frameworks, with structural quality pertaining to how

balanced the argument is and in particular, the use of rebuttals and counter claims. Upon partitioning the analysis of arguments in this way, it is clear that there are a number of approaches and levels from which to determine argument quality.

The considerations that need to be incorporated into any approach for argument quality analysis each represent a unique research domain. For example, the construction of an argument using available knowledge via interaction with another could have the potential to increase argument skill competency. In turn, this competency can influence critical thinking ability and the quality of discourse with another. Herein lays the difficulty in developing a unified argument model that accounts for all the constraints mentioned. Of course, to complicate things further there are a myriad of other factors not discussed or presented here that could also potentially influence argument structure. Just two examples of these additional factors may be the current affective state of the arguer, or the perceived knowledge and assumptions made about the receiver. The quality analysis approaches that this thesis will adopt propose that rationales that incorporate a two sided approach are of greater quality than those that do not. The adoption of a variety of argument analysis methods should enable an evaluation of these approaches and the development of improved quality analysis frameworks.

The argument and explanation research has also revealed that argument can vary as a function of several constraints. One of these contextual constraints will be addressed in this thesis, that of the perception of direction held by the author, either as self or other directed, of an argument and the impact of this on the strategies adopted and resulting argument quality. The impact that the manipulation of this perception may have on the argument quality and structure is not clear, however, research examining the self-explanation effect has potentially shown how the perception of direction could impact upon argument quality, particularly in terms of structure. Contrary to previous research the methodology adopted in the thesis will focus on manipulating the perception of direction (self or other) whilst keeping the actual interaction consistent to examine whether variations in this perception can produce measurable differences in rationale structure and decision confidence.

Rationale style arguments will be the focus of investigation, as these appear to be a very common type of elicited argument in research and decision making. Rationales are also most often intended to be complete arguments that do not necessarily require interaction with another to be fully realised. They are considered both an argumentative and a reflective, explanatory style of thought and have been defined in the previous research as an

"explanation of reasons underlying ones decisions, conclusions and interpretations." (Xiao, 2013b, p. 524). This multifaceted purpose and content makes the rationale a potentially rich source of varied argumentative structures. In addition, rationale style arguments are investigated in many domains, including design and education, thus the scope of utility for the findings should be wide.

Part Three: Empirical Work

This third part of the thesis will describe, in chronological order, the entire body of experimental work that has been carried out. The work has been conducted with a view to address the first five research questions and to provide extensive data that will be utilised in investigating the remaining four. The discussion of the relevant research conducted previously in part two, informs the choice of frameworks that can be used to effectively analyse the arguments elicited and provides a theoretical basis for the findings. Rhetorical Structure Theory and the derived automated text approaches are prevalent methodological tools across the experimental work and as such the fourth part of the thesis will seek to evaluate these methods for the purpose of argument analysis and suggest possible adaptations.

The focus of the experimental work will be the construction of rationale style arguments and primarily the perception of intended rationale direction and possible future use as a determinant of argument structure and decision confidence. The literature explored in the first part of the thesis demonstrates how useful these types of arguments may be in terms of studying argumentative approaches and decision making. The intended direction of a rationale as either self or other as a pertinent factor in argument analysis is discussed in the final section of chapter 6 and highlights the need for explicitly examining the perceptions that authors hold when constructing arguments in combination with the physical context.

5 Self versus Other Directed Rationales: Comparing Confidence and Content

5.1 Introduction

An exploratory study was conducted to gain an initial insight into the basic effect of perceived rationale direction in terms of the differences in structures within self and other directed rationales and the confidence held in a decision.

The study will begin to examine the questions prompted from the interaction hypothesis of whether people are motivated to construct different arguments in response to the mere prospect of others interacting with it. Unlike previous research, the rationales will not be elicited aloud in order to minimise feedback and concerns about immediate scrutiny from others. The perception of direction will be influenced by a prompt informing the authors that their rationales will be shared and used by others.

The previous research by Koriat, et al (1980), Sieck & Yates (1997) and Lu, et al (2011) suggest that confidence can be enhanced as a result of lengthier explanations and a more justificatory approach. Therefore it will be considered more likely that those in an other directed prompt group will construct more complex rationales and that an increase in length will be linked to an increase in confidence.

In order to guide the rationale construction of the participants, the DSA framework elements of Questions, Options and Criteria, discussed previously, will be adapted as a prompt to encourage the production of a full, relevant and argument based rationale. The study evaluates the use of this particular style of rationale prompting and the use of a prompt to cue participants to construct either self or other directed rationales without direct interaction. A content and structure analysis of the rationales will also be conducted to ascertain any differences in reasoning styles between the groups.

5.1.1 Research Questions

From the initial examination of the literature the following three research questions were generated:

1. Does the perception of direction (either self or other directed) held by an author when constructing a rationale style argument influence perceived decision confidence?

2. Does the perception of direction (either self or other directed) and future use held by an author when constructing a rationale style argument influence the structures within and the quality of arguments?

3. Does the length and structure of the rationale style argument have any bearing on perceived confidence in a decision based on the rationale?

5.1.2 Hypotheses

The following hypotheses were constructed to investigate the research questions:

H1: The prompt for future use and sharing of a written rationale will result in a higher level of perceived confidence in a decision compared to those who have been informed that the rationales will not be used or shared.

H2: The prompt for future use and sharing of a written rationale will result in more complex rationales (in terms of the number of 'Options' elements observed). The 'Options' element from the Design Space Analysis framework represents a consideration of alternatives and thus a wider processing of the information available.

H3: The prompt for future use and sharing of a written rationale will result in rationales that have a more complex rhetorical structure (as measured by the PDTB parser). Those in the 'Other Directed' group will generate more 'Argue' type relations within the rationales, compared to those in the Self Directed group.

H4: There will be no significant difference in the use of 'Analyse' or 'State' type relations between the Other and Self Directed groups.

H5: The length of the rationale will have a positive relationship with the level of perceived confidence in a decision based on the rationale.

5.2 Method

5.2.1 Participants

The other directed (OD) group comprised of 17 participants. A total of 13 participants comprised the self-directed (SD) group. Participants were aged between 18-25 years, with 18 females and 12 males. The sample was selected from undergraduate students at the University of Bath, on the basis of availability in response to a mail shot.

5.2.2 Design

5.2.2.1 Independent variables:

A between subjects design was used with the direction of prompt as the independent variable with two levels:

1. **Self directed prompt**: prior to constructing the rationales participants will be informed that their rationales will be kept private.

2. **Other directed prompt**: prior to constructing the rationales participants will be informed that their rationales will be made available to others to assist in their decision making.

5.2.2.2 Dependent Variables:

The full list of dependent variables can be seen in Table 9. A full explanation of the analysis methods for these variables can be found in the procedure section.

	Dependent Variable	Specific Measurement
1	Decision Confidence	Expressed as a percentage
2	DSA elements used	Number of Question elements
3	Rationale Word Length	Mean rationale length
4	Argue type relations	Number of Concession, Alternative and Contrast relations.
5	State Type relations	Number of Conjunction, Restatement and Instantiation Relations
6	Analyse Type Relations	Number of Asynchronous, Synchronous or Cause relations

Table 9 Summary of dependent variables: attitude, structure and quality

5.2.3 Procedure

5.2.3.1 Directional Prompting

Prior to constructing the rationales the participants were informed, via a written prompt, that their rationales would either be kept private (self-directed) or be made available to others to assist in their decision making (other directed). The wording of the prompt was designed to strengthen the perception of direction with the suggestion of future use by another.

The aim of the prompt was to elicit a sense of writing an argument for the use of another in the future. This was considered a more powerful prompt for altering the perception of direction, without actual physical interaction with another, than simply stating 'Please write with another person in mind'. The suggestion of future use for the argument should prompt a sense of writing for another.

5.2.3.2 Decision Task

All participants carried out the task individually using pen and paper, in a room shared with the experimenter. The task consisted of a fictional decision scenario, requiring the participant to imagine they were in charge of a drug trial. The use of a novel and fictional topic as a scenario narrows information use to the resources given, reducing the scope of

possible variables that could impact performance. Each participant was required to decide which fictional medication to prescribe to treat two patients presenting symptoms of an illness. The task brief was presented (see Appendix 5) to each of the participants who were instructed to write their answer and produce a written rationale.

The brief contained isolated descriptions of each patient with information pertaining to age, family life and severity of symptoms along with expected prognosis for treatment. Each drug was allocated a score of effectiveness in treatment of the strains of the illness, expected side effects, severity of side effects and likelihood of side effects. These criteria were designed to act as trade-offs, of which the participants could make use to decide a 'best fit' solution for each patient. The participants were required to use the information provided to construct an argument for their choice. The restrictive nature of the task used would theoretically validate direct comparisons of qualitative features of the rationale and the word count, as it could be considered the content was more likely to be of similar type of information and contain the same types of trade-offs and knowledge.

In terms of the ordering of task and choice, participants were initially asked to write a rationale and then confirm their choice on the same page as the rationale. They were required to do this for both patients; therefore each participant will essentially produce arguments containing two rationales upon completing the task. The rationales were prompted using a loose QOC format, asking them to state (as a suggestion) what questions they may have asked themselves when coming to their decision, what other options they considered and what criteria was most important to them when considering their choice. The QOC framework was used as a prompt to help participants be more aware of what a rationale should look like, as opposed to relying on an individual interpretation of a 'rationale' which may not result in comparable responses for analysis.

5.2.3.3 Confidence Measures

To address the first hypothesis, regarding whether prompting for a future use of a written rationale impacts the perceived level of confidence, upon completion of the task all participants were asked to estimate using a percentage, how confident they felt they were that they had made the correct decisions for the task. The reasoning behind using confidence as a dependent variable lies in the concept that confidence can act as a predictor of task success (Feather, 1968) and participants have been shown to actually approach a task differently as a result of varying confidence from prior experience.

Confidence allows quick measurement and also allows for participants to express uncertainty or indeed ambivalence towards a particular choice. However, it must be noted that confidence is not the only factor that indicates the attitude towards a decision (Heath and Gonzalez, 1995), but it still may be useful way to make explicit the more affective aspects of the decision.

5.2.3.4 Analysis Methods

To address the remaining hypotheses concerned with rationale structure, the rationales will be analysed by a single analyst using a number of approaches. The approaches are described in detail below.

5.2.3.4.1 Design Space Analysis Framework

To address the second hypothesis the rationales were analysed on a sentence level basis to identify the basic purpose of each. The basic purpose of each statement would be categorised according to whether it offered an 'Option' for the decision solution, a description of the key 'Criteria' for the decision or a 'Question' which was pertinent to the decision. The Criteria aspect refers to statements that list elements of the material that were central to the decision. The Questions aspect refers to key questions that the author had considered prior to reaching a decision. The Options aspect of the QOC framework pertains to the alternatives that were considered and justifications for the rejection of these in favour of the chosen solution.

The Options aspect of the QOC framework also appears to be more complex in nature than the Criteria and Questions aspects of the framework, as this element requires evaluative and justificatory processes when considering alternatives and perhaps a wider consideration of the material presented at the time of rationale construction. It was also considered appropriate to analyse the rationales using the QOC framework to ascertain whether the framework was adhered to and did in fact gave rise to rich and effective rationales.

5.2.3.4.2 PDTB Parser Analysis

To address the third and fourth hypotheses concerned with the use of relation categories between the groups, the rationales were compared using an automated PDTB parser

analysis. This was also carried out to ascertain the viability and efficiency of using an automated parser to analyse rationale style arguments. The PDTB parser identifies argumentative structures primarily by categorising the discourse connectors present in the text that signal coherence and give meaning. This analysis will give a basic indication of some of the argumentative type relations present. The PDTB parser includes the labelling of explicit and implicitly signalled relations but as the implicit and explicit distinction between the relations is not a feature of specific interest to the analysis; the relations are considered equal for the purposes of this investigation.

The reasoning styles within the rationales will be considered in terms of two dimensions, either 'Argue' or 'State' types. The categories are extracted and adapted from the work by Mentis et al. (2009) using Rhetorical Structure Theory (see section 3.1.3.3 for discussion). The categories were determined based on the similarity of the definitions and the purpose of the relations defined in the original research. The State category of relations refers to those elements that offer information to support an argument without any interpretation. The Analyse category includes those relations that offer insight into the effectiveness of the data included, such as whether the author has stated that evidence is 'strong,' or is indicative of something else. The final category of Argue includes those relations that may be intended to have a persuasive effect on the reader and also indicate that alternatives have been considered. The Argue type relations may relate to the use of rebuttals, the most complex type of argument according to Kuhn (1991). These categories are used to guide the hypotheses and structural analyses. The possible correspondence of the rhetorical relations with the Toulmin model is discussed in section 10.2.4. A summary of PDTB parser relation groups can be seen in Table 10.

Relation Category	PDTB Relations
'State' Relations	Conjunction
	Restatement
	Instantiation
'Analyse' Relations	Asynchronous
	Synchronous
	Cause
'Argue' Relations	Concession
	Alternative
	Contrast

Table 10 Relation categories for the PDTB parser argument analysis.

In the original paper, Mentis et al. (2009) used the categorisation of relations to analyse discussion in a collaborative argument context, the following investigation is fundamentally an individual context and therefore the relations that do not fit in the predefined categories will still be identified but will not be included in the formal hypotheses. The Contrast relation was included in the Argue category as in the PDTB parser label definitions (see Appendix 4), this relation forms part of the Comparison class along with Concession. Therefore it would be appropriate to treat this relation as comparable to the Concession relation and label it as having an argument purpose.

For the Analyse category of relations, no significant differences were found in the previous research using this category between more reflective argument and other directed conversation (Mentis et al., 2009). This suggests that the use of Analyse type relations will not differ depending on rationale direction. The previous research also revealed that the use of the State relations were consistent across less interactive and fully interactive discussion. Therefore, the State type of relations - similar to the Analyse group - will be considered likely to have a similar distribution across the groups as these represent the use of backing and data.

To account for those relations that may be detected but that are not included in the three categories, a fourth category of 'Additional' relations will be added when reporting the findings (see Table 16 in the findings for examples). The relations in this category are not

thought to have an argumentative function and will therefore be excluded from analysis in this chapter.

5.2.3.4.3 Rationale Length and Confidence Analysis

To address the final hypothesis that concerns whether rationale length is linked to the level of perceived confidence in a decision, the total word length for each individual rationale in both combined groups will be correlated with the corresponding confidence ratings. In addition, the rationales will be grouped according to length and an analysis between the groups of the confidence ratings will be carried out.

5.3 Findings

The initial findings for task choice will be considered followed by a summary of findings relevant to each hypothesis.

5.3.1 Task Choice

The summary of choices made for patient A and B can be seen in Table 11. The majority of participants in the SD group opted to use drug one to treat patient A (followed by drug three) and chose drug three to treat patient B (closely followed by drug two). In comparison, the majority of participants in the OD group chose drug three to treat patient A (followed by drug two) and drug two to treat patient B (closely followed by drug three). Drug one was the moderate choice in terms of side effects and effectiveness, whereas drug two, by contrast, was much more effective but more severe in terms of side effects. Drug three, which was a popular choice for both patients, represented a 'safe' option with the least effective effects but the mildest side effects.

	Self Directed			**Other Directed**		
	Drug One	Drug Two	Drug Three	Drug One	Drug Two	Drug Three
Patient A	9	1	3	7	1	9
Patient B	1	5	7	0	10	7

Table 11 Summary of drug choice for each patient in the OD and SD groups.

Due to the uncertain nature of the task, it is not surprising that many opted for the 'safe' option of drug three. There was no significant difference observed between the choices of drug made in both groups for patient A (χ^2 (2) = 2.766, p = .251) or patient B (χ^2 (2) = 2.172, p = .338).

5.3.2 Content and Structure Analysis

A sample of the rationales from both of the experimental groups is available in Appendix 6.

5.3.2.1 Rationale Length

The average rationale length for the SD and OD groups was examined and the findings summarised in Table 12. Again, no significant difference was found (U = 332.5, Z = -1.634, p = .102) between the groups, although the mean rationale length in the OD group appears slightly higher.

	N	**Rationale Length (M)**	***SD***	**Median**
Self-Directed	26	35.19	18.43	33
Other Directed	34	43.79	22.15	40

Table 12 Comparison of means for rationale length between groups.

Although not a statistically significant difference, in this small sample it gives an encouraging indication that those in the OD group constructed rationales that were longer in length than those in the SD group. It would be interesting to uncover why this might be in a future study by conducting a more in-depth content and structure analysis. Previous research has suggested that longer length arguments may be a basic indication of improved decision performance and more cognitive complex processes.

5.3.2.2 Hypothesis 1: Confidence Comparison Between Groups

To address the first hypothesis that those who are prompted for future use as other directed will have a higher perceived confidence in their decision a Mann Whitney U analysis was performed.

	Confidence Ratings		
Group	**Mean (%)**	**SD**	**Median**
Self Directed	72.69	8.57	75
Other Directed	73.53	14.93	77

Table 13 Comparison of average confidence percentage ratings for the SD and OD group.

From observing the patterns of confidence (see Table 13) ratings very few, as expected, fell below the 50% mark, as it would be assumed they would have simply made an alternative decision. However this is not entirely straightforward as there were three drugs to choose from to treat two patients and it was not simply an 'either or' situation. No significant difference was found between the mean confidence ratings for the OD and SD groups (U = 378, Z = -.964, p = .335).

5.3.2.3 Hypothesis 2: Use of Options Comparison Between Groups

To address the second hypothesis that those who are prompted for future use and are other directed will construct more complex rationales in terms of the number of 'Options' elements observed, a Mann Whitney U analysis was performed on the data. Examples of rationale segments identified for each DSA category are presented in Table 14.

Category	Rationale Segment Example
Question	"Will they benefit from the drug treatment?"
Option	"The side effects are moderate (…) which would make drug 2 too dangerous."
Criteria	"The side effects were important as was the likelihood of success"

Table 14 QOC analysis example statements

The means per group of the QOC components identified within the rationales are compared, and visually summarised in Figure 8. The descriptive statistics for each group are presented in Table 15.

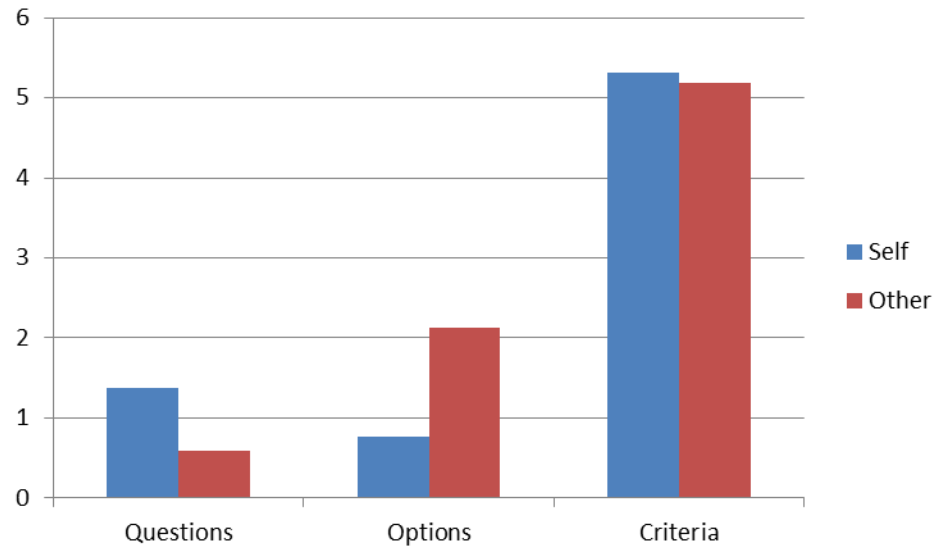

Figure 8 Means (y-axis) for individual QOC (x-axis) elements per group.

As can be seen in Table 15 the OD group contained on average considerably more Options type components (M=2.11) than the SD group (M=0.77). This difference was revealed to be significant following a Mann Whitney U statistical test (U=59.5, Z= -2.201, p=.028, d = .93).

DSA Element	Group	Mean	SD	Median
Question	Self Directed	1.38	2.84	0
	Other Directed	.59	1.33	0
Option	Self Directed	0.77	0.83	1
	Other Directed	2.12	1.87	2
Criteria	Self Directed	5.31	2.36	1
	Other Directed	5.18	2.13	6

Table 15 Summary of average DSA Questions, Options and Criteria elements in the SD and OD group.

A further post-hoc analysis was conducted to compare the number of Criteria (U = 108, p = .934, Z = -.106) and Question (U = 89.5, p = .385, Z = -1.130) elements and no significant differences were found between the groups.

5.3.2.4 Hypothesis 3: Argue Relations Comparison Between Groups

The third hypothesis states that the prompt for future use and sharing of a rationale will result in a more complex rhetorical structure (in terms of the use of Argue relations) than those who are prompted for self directed rationales. To address this, the rationales were analysed using the automated web based PDTB parser and the relations found grouped according to the categories discussed in the procedure section. Spelling and punctuation errors within the rationales were corrected prior to being analysed by the parser to reduce errors based on these features.

The descriptive statistics for each type of relation identified using the PDTB parser within both groups can be seen in Table 16. The medians for each relation are reported for completeness and where non-parametric analyses are used. However, due to the high number of zero values in the data the medians are less informative than the means. Therefore, for visual representation of the data, the mean value will be used.

		Self			**Other**		
Category	**PTDB Relation**	**Mean**	**SD**	**Median**	**Mean**	**SD**	**Median**
'State'	Conjunction	0.69	1.12	0	0.59	0.78	0
	Restatement	0.08	0.39	0	0.24	0.65	0
	Instantiation	0.00	0.00	0	0.03	0.17	0
'Analyse'	Asynchronous	0.15	0.37	0	0.00	0.00	0
	Synchronous	0.12	0.33	0	0.18	0.39	0
	Cause	0.92	1.06	1	1.12	1.17	1
'Argue'	Concession	0.12	0.33	0	0.35	0.54	0
	Alternative	0.00	0.00	0	0.03	0.17	0
	Contrast	0.12	0.33	0	0.26	0.57	0
Additional	Condition	0.08	0.27	0	0.21	0.41	0
	Ent Relation	0.27	0.53	0	0.24	0.50	0

Table 16 Summary of all PDTB Parser relations identified in the rationales for the SD and OD groups.

The category of Argue relations was considered to include Concession, Alternative and Contrast. As the SD group rationales did not contain any instances of Alternative relations

these were excluded from further statistical analyses. The differences between the use of the Argue type relations in the OD and SD groups are apparent in Figure 9. It appears that the presence of Concession and Contrast relations is higher in the OD group, which is perhaps indicative of more complex reasoning styles and a more balanced approach to argument in this group.

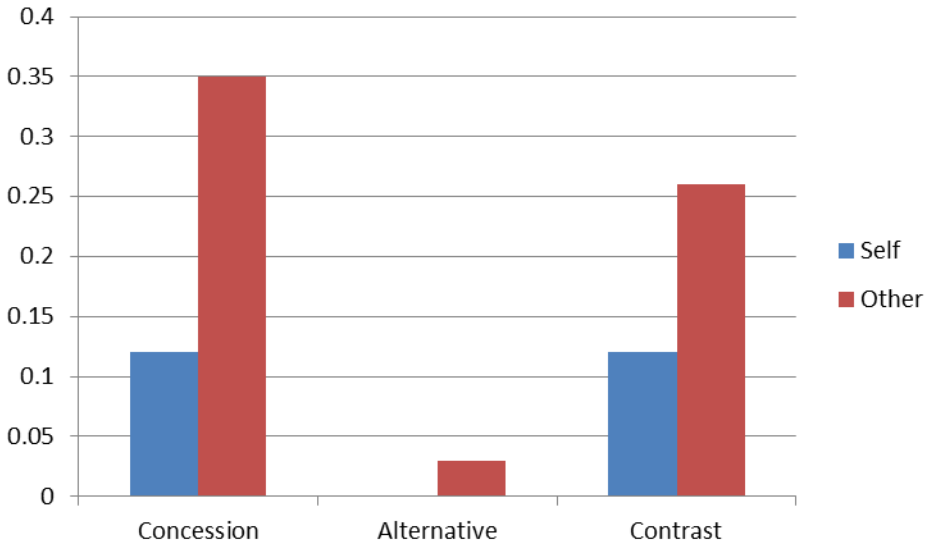

Figure 9 Mean (y-axis) PDTB 'Argue' relations (x-axis) within each group.

A significant difference was found between the number of Contrast relations identified, with those in the OD group producing rationales with significantly more Contrast relations than those in the SD group ($F= 3.304$, $t(28)= 2.305$ $p= .029$, $d = .30$). However, this variable did not pass the homogeneity of variance test and the results were not significant when reanalysed using a Mann Whitney U test. No significant difference was found between the use of the Concession relations between the groups ($U = 86$, $p = .218$, $Z = -1.233$).

5.3.2.5 Hypothesis 4: Comparison of Analyse and State Relations Between Groups

The fourth hypothesis stated that the OD and SD groups would not differ in the preference for State type relations. In order to address this, the Conjunction and Restatement relations were compared between groups. The distribution of these relations between the groups can be seen in Figure 10. The Instantiation relation was removed from analysis as this relation was not present in the SD group rationales. No significant differences were found for either

Conjunction (U = 437, p = .932, Z = -.085) or Restatement (U = 395, p = .178, Z = -1.347) relations between the groups.

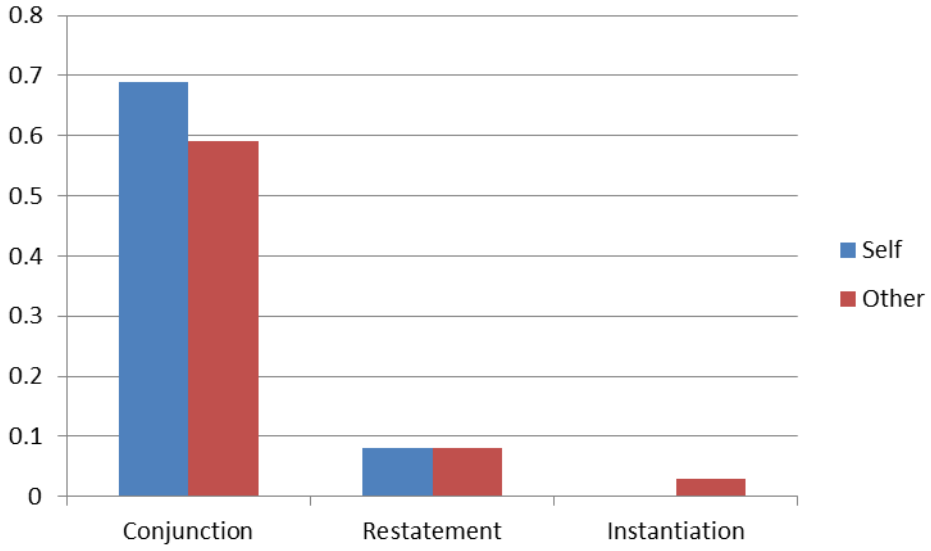

Figure 10 Mean (y-axis) PDTB 'State' relations (x-axis) per group.

The fourth hypothesis also stated that there will be no significant difference in the use of the Analyse type of relations between the groups. The distribution of these relations between the groups can be seen in Figure 11.

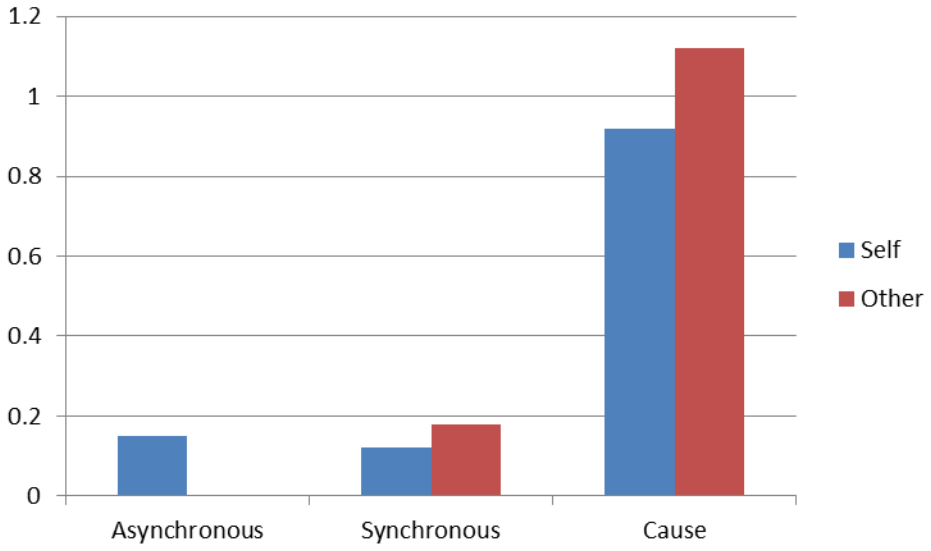

Figure 11 Mean (y-axis) PDTB 'Analyse' relations (x-axis) per group.

The Asynchronous relation was not present in the OD group rationales and therefore was excluded from statistical analysis. No significant difference was found between the groups for the Synchronous (U = 415, p = .515, Z = -.651) or Cause (U = 403, p = .535, Z = -.651) relations.

5.3.2.6 Hypothesis 5: Relationship Between Rationale Length and Confidence

The final hypothesis was concerned with the relationship between the length of the rationale and the perceived decision confidence and it was predicted that there will be a positive relationship between these two variables. In order to analyse this, the word count data for all of the rationales was combined for both groups to statistically examine the relationship between rationale length and confidence scores.

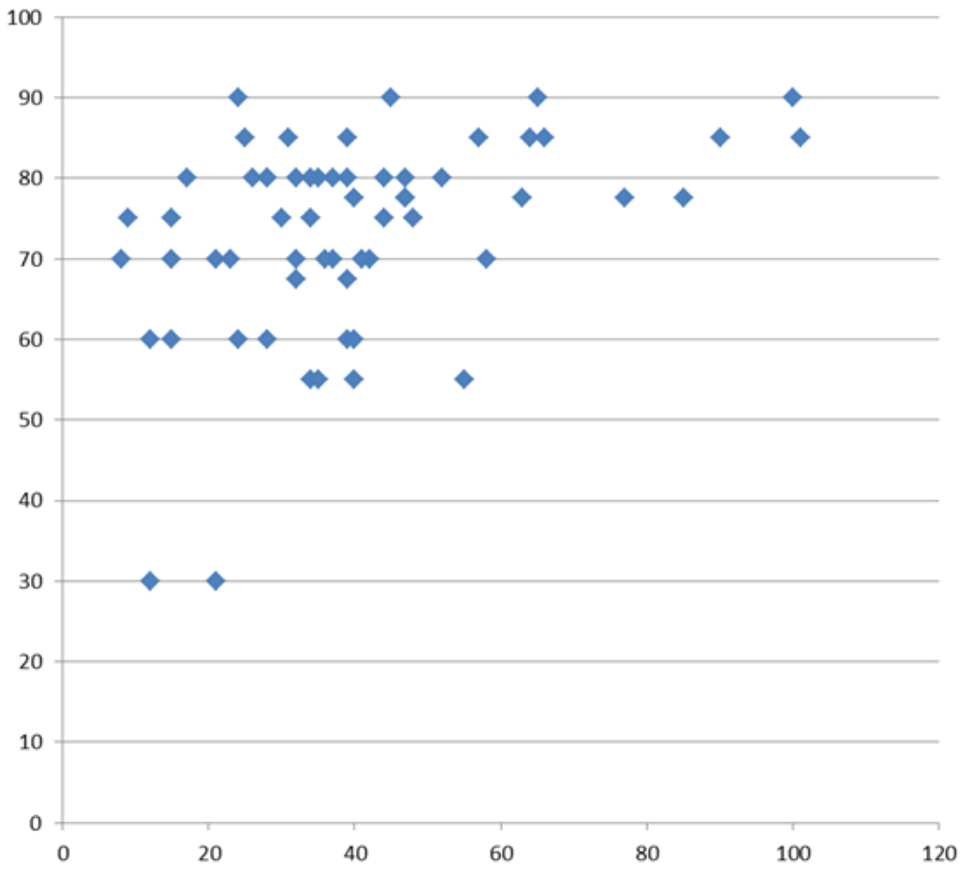

Figure 12 Word count (x-axis) and confidence percentage ratings (y-axis)

A Spearman correlational analysis was also conducted to ascertain the strength of this trend and a statistically significant positive correlation between the total word count and confidence scores was found ($r_s(60) = .384$, $p<.01$). As shown in Figure ***12*** the trend seems to be that the longer the rationale is in length, the higher the average confidence rating seems to be.

Group	**N**	**Word Count Range**	**Mean**	***SD***	**Confidence Mean**	***SD***
1	14	1-25 words	17.21	5.79	66.07	17.67
2	24	26-40 words	34.38	4.36	71.98	9.89
3	13	41-60 words	48.23	5.59	75.77	8.56
4	9	60-101 words	79.00	15.54	83.61	5.02

Table 17 Descriptive data for the separated word count group

In order to examine this relationship further, the rationales were categorised according to length and then grouped for analysis as shown in Table 17. The mean confidence scores between each of the word count groups were compared using an independent samples Kruskal-Wallis analysis and a significant difference found ($\chi2(3) = 12.411$, $p = .006$). Pairwise comparisons between the word count groups revealed that Group One (lowest word count) and Four (highest word count) were found to have a significant difference in confidence scores ($Z = -3.353$, $p = .001$, adjusted sig. = .005.) A significant difference was also found between the confidence scores in Group Two and Four ($Z = -2.951$, $p = .003$, adjusted sig. = .019).

5.4 Discussion

The findings support the previous research to some extent and prompt further investigations into the impact of directional prompts on rationale construction and decision confidence. There is an evident need to further clarify whether the differences observed in rationale structure arise as a result of a perceived self or other directed approach and it is thus necessary to examine how these structures may relate to rationale quality, task performance (in terms of information recall) and decision confidence.

5.4.1 Hypothesis 1: Confidence Comparison Between Groups

The null hypothesis is accepted in this case as no statistically significant difference was found for the confidence scores between the groups. Although not statistically significant, the higher mean confidence for the OD rationale group is interesting and could be a result of a number of factors. There are three possible considerations for confidence effects proposed which are applicable to rationale elicitation tasks. The first being the 'effort-performance belief' effect, whereby a person anticipates their performance level as in part as an increasing function of how much effort they put into the task (Sieck &Yates, 1997). Confidence may also be bolstered as a result of the rationale construction process inducing some sort of learning. For example if given a free recall task it may be found that participants who constructed their rationales as other directed may recall more task details than those who constructed self directed rationales. If a sense of greater knowledge is achieved as a result of increased attention to the task this may induce a feeling that a better decision has been made.

Koehler (1991) demonstrated that constructing an explanation for a potential occurrence actually increases the belief in the likelihood of that occurrence. This effect can occur as a result of the process of assembling a coherent argument for one position, as one may tend to focus solely on gathering information to support that particular position and ignore negative or contradictory evidence.

It may also be that those who constructed more balanced and coherent rationales had a greater sense of understanding of the task as a whole and felt confident that they had assessed all possible options adequately. The more coherent and balanced rationales appeared more frequently in the other directed group. It may be that the perception of an other directed approach prompts a different strategy to fully justify and counter possible criticisms. This strategy may not be so pertinent in the minds of the authors who know their rationales will remain unscrutinised by others. This confidence trend will be examined further in the next investigation.

5.4.2 Hypothesis 2: Use of Options Comparison Between Groups

This hypothesis is supported with a statistically significant difference found between the two groups. The OD group appeared to construct rationales with a higher number of

Options elements. This is intriguing as the Options aspect of the rationale could be considered to indicate to a reader that a balanced approach was taken, and that the author has considered all sides of the argument. The externalisation of this may be triggered if an author feels that their decision is to be scrutinized and this strategy of considering options enables counter arguments to be dismissed before they are made.

An explanation for this observation may be that participants opted out of mentioning Options in their arguments as the available solutions of which drugs to choose were already mentioned in the brief and therefore considered a given. Thus, the restatement of the options may have been considered a laborious and pointless aspect in terms of their argument. In the case of a more open task, the 'other options' may not be explicitly obvious so may be more worthy of note.

The consideration of Options within a rationale may be an indicator of argument balance, in that the author wishes to convey that all possible alternatives have been considered and in light of this, the most plausible option was selected. This tendency to incorporate a more balanced argument appears to be favoured by those in the OD group.

The use of the QOC framework as a prompt for rationale elicitation largely resulted in disjointed rationales that were often only focused on one particular aspect of the QOC framework such as Questions, and were not complete or coherent argumentative pieces. This suggests the use of the QOC framework as a prompt may have impacted on the fluid nature of rationale construction and presented a further cognitive load alongside the already novel and unfamiliar task scenario.

5.4.3 Hypothesis 3: Argue Relations Comparison Between Groups

This hypothesis is not fully supported by the statistical analyses. The OD group were found to contain rationales that had a higher number of 'Argue' relations in terms of Contrasts. The difference between the groups was found to be significantly significant using an independent t-test analysis, however this was not the case when analysed using a Mann Whitney U test. This does indicate a possible trend that those in the OD group appear to use more relations that pertain to balance and the contrasting of ideas. This may be a result of participants responding to the Options prompt in the QOC based brief which would inevitably include contrasting of alternatives. It appears that those in OD group may display

an increased tendency to demonstrate that all possible options had been considered. Again, this is a trend that will be investigation further with a more in depth analysis in the next chapter.

5.4.4 Hypothesis 4: Comparison of Analyse and State Relations Between Groups

This hypothesis was supported as no significant differences were found between the groups in terms of Analyse or State type relations. This suggests that the Analyse and State type relations are equally prevalent regardless of perceived direction and supports the findings by Mentis et al (2009). This finding suggests that the Analyse and State relations do not vary substantially as a result of directional prompting and it would therefore be appropriate to exclude these variables from further investigation.

5.4.5 Hypothesis 5: Relationship Between Rationale Length and Confidence

This hypothesis was supported by a statistically significant positive correlation being found between the confidence scores and the rationale length. A possible mechanism behind this trend may be that the construction of a rationale enables participants to feel confident that the options have been effectively evaluated and justified. This could be an example of the effort-performance belief effect (Sieck and Yates, 1997). However, it could be argued that those participants who had higher confidence in the task took the task more seriously and therefore constructed lengthier rationales, therefore it is not clear whether constructing a lengthier rationale is a determinant of confidence.

5.4.6 Limitations

The apparent marked differences in word count between the OD and SD group needs to be confirmed with a much larger sample and with a view to fully investigate the structures and argumentative elements within the rationales produced. The mean lengths of the rationales within both groups are somewhat small for an in-depth analysis of structure and content. This could be mediated in a future study with a more familiar task scenario and a more open rationale prompt.

A number of methodological issues were highlighted in this exploratory study, which may influence the findings and will inform the approach of subsequent research. Firstly, the only information available to the participants that they could use to reach a solution was provided in the task brief. Prior knowledge would not be applicable as the situation was entirely fictional and novel. However this narrowing of scope considerably reduces the real world validity as most uncertain decisions are approached and evaluated using intuition and prior knowledge in combination with the available external resources. The novel and fictitious aspect of the task may also have reduced confidence and thus inhibited rationale construction as participants were not familiar with the topic.

However, having seen these trends in confidence, rationale length and the OD versus SD effect, it is not possible to draw concrete inferences on whether the construction of a rationale itself was an indisputable factor in these results, as the order of choice and rationale elicitation was not carefully controlled and decisions could have been made before or after the rationale was constructed. If inclined participants could make their choices prior to constructing a rationale, perhaps inhibiting the externalisation, as the participant is more focused on the decision as the task goal and not the production of an argument. The exact ordering and impact of rationale within a decision needs to be controlled and this is a factor that will be considered in the next investigation. In this research, the externalisation of a rationale needs to be an active 'part' of the decision process and not simply an ad hoc structuring of ideas post decision, as research has shown that information seeking behaviour is halted if a decision is already reached. In addition, the use of the QOC framework as a prompt for rationale elicitation may have been too prescriptive in nature, and participants may have been simply answering the questions as opposed to concentrating on creating a valid and coherent argument.

In addition, the use of the word 'correct' when prompting participants for their confidence scores implied that there was a correct solution as opposed to a best fit or most appropriate route. As a result of this, confidence ratings may have been reduced at the time of posing the question as no 'correct' solution was obvious in the task. Interestingly, none of the participants reported that they were 100% confident in their decision which may have been as a result of this unease. In future iterations, in order to assess true confidence in the solution and rationale that it is not attenuated negatively by the use of such restrictive terms as 'correct', the phrase 'best decision' may be used, as this implies that any solution could be justifiable.

Further to this, the task itself may have posed a difficulty to the participants and interfered with the cognitive processes in terms of the choice response as no obvious solution could be seen even after careful examination of the materials. The task was designed to narrow the scope of the resources available to construct an argument to ensure that the resulting rationales were comparable to each other in terms of content. However, upon examining the data only very few of the respondents actually constructed a full QOC style rationale. Most participants simply used a justificatory approach and generally failed to structure responses in the QOC style. This novel and fictional task coupled with the QOC style prompting for the rationales may have inhibited the externalisation of the arguments. Inhibition may have occurred from over structuring and therefore increased cognitive load by presenting prompts that required excessive processing. Some participants may also have been reluctant to comment extensively on a novel topic. To attenuate this, a more open prompt and potentially accessible topic will be implemented in future work. The OD and SD prompts used in the study, to facilitate the perceived direction of the decision, appeared to produce small but measurable effects. However, the addition of a perception check measure would be needed in a future study to ascertain whether the participants actively considered their rationale as other or self-directed, so the strength of the impact of the prompts could be ascertained. This will provide insight into whether participants are aware of a conscious decision to construct a self or other directed argument.

5.5 Modelling the Findings

The trends observed from this initial exploratory study were used to inform a model of the processes that impact upon rationale style argument structure and confidence held in the decision. The model depicted in Figure 13 scopes the processes and the considerations brought to light by the findings. The model tentatively demonstrates, using some of the QOC elements, how argument structure may be influenced by the perception of future interaction (or the possibility of) with an externalised argument. The nodes represent broad internal and external aspects.

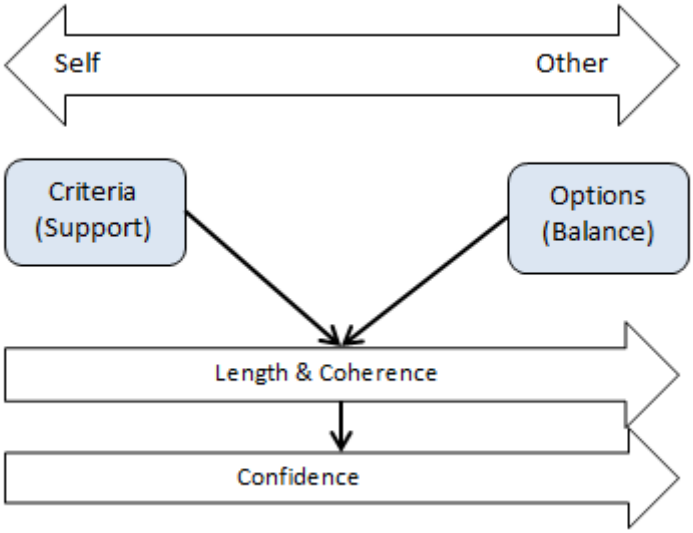

Figure 13 Initial model of the perceived direction effect

The 'self and other' arrow bar at the top of the model represents the perceived direction. The arrows for 'Length & Coherence' and 'Confidence' indicate that these aspects increase as a result of an intended other directed approach by the author. The linking arrows between model elements represent apparent positive relationships between them. For example, those rationales that contain more elements within them are possibly going to be more coherent and lengthier and this appears to have a positive relationship with the perceived confidence although this is not considered causal in nature.

Although the participants were manipulated to view their rationales as either totally private or to be shared with others, the actual interactivity level for both groups was private. So far the findings have suggested that the perception of the possibility of future interaction with an externalised rationale appears to shift the argumentative strategy to mimic that of direct interaction with another. That is, the arguments become more coherent, balanced and also lengthier. The length of the rationale may be indicative of an extended use of more complex argumentative strategies. In addition the perception of self or other directed rationales appears to impact upon confidence, with those in the OD group who constructed lengthier and more complex rationales, rating their confidence in a decision as higher than those who constructed less complex rationales.

The findings do indicate some trends worth considering for the prompting of direction when constructing an argument. It would be worthwhile to add additional detail to the model with future work by indicating if the different argumentative strategies by each group lead to a deeper engagement with task material. It would be informative to assess the

difference in approaches using an extended number of argument analysis frameworks to enable a more fine grained argument structure to be presented in the model. This could be verified with pre and post testing for task information recall. If this could be ascertained it may give an indication of the internal cognitive processes that differentiate the two approaches and how each group is interacting with the task material. This has implications for findings within the Self explanation literature in that it suggests that the 'self explanation effect' may be in fact be influenced by a perception of writing for others.

The model helps to visualise the impact of perceived direction of rationale on an externalised argument, independent of actual interactivity level at the time of argument construction. This perception changes the actual strategy from thinking you have written for others (which many may think they do automatically) to actually doing so and behaving in manner which is more prevalent in direct interaction. The two sides of the model are not prescriptive or exclusive for argument structure, and both levels of interaction, either perceived or actual can facilitate any type of argument structure. However, the model shows the proportions of argument types and elements that are more likely to persist in each context depending on the perception held by the author and thus different strategies could be the culprit.

5.6 Next Steps

The results have indicated a number of interesting trends. The perceived direction of the rationale (to self or others), even in a non-interactive physical context appears to have an impact on decision confidence ratings and rationale structure.

In order to more reliably examine any differences in confidence or rationale structuring in response to self or other directed prompting, it would be useful to assess whether these prompts were attended to by asking participants themselves to rate the intended direction of the arguments.

The rationales elicited will need to be more complex, coherent and lengthier than the rationales produced for the exploratory study. This will enable a more complex structural analysis of the arguments. Additionally, the question of whether the production of rationales helps with information retention and subsequent recall will be examined using pre and post testing. The use of the automated parsers will also be investigated in more depth with a

larger sample, given the indication in this study of a possible difference between groups regarding rhetorical argumentative structure and reasoning styles.

6 Self versus Other Directed Rationales: A Comparison of Reasoning Styles

6.1 Introduction

The findings from the first study suggest that the perception of rationale purpose may impacts on argument quality and confidence to some extent, but the question of whether this impact is indicated or measurable in the externalised rationales or explanations produced has not been fully addressed.

The use of explanation in open domains has only been tentatively approached in the literature, possibly due to the wide ranging nature of knowledge types and skills required in these less structured topics. These types of domains invite an argumentative approach to enquiry, as many are controversial and uncertain in nature.

For the purposes of this study, a decision scenario was required that represented a relatively open ended domain, that most people may be familiar with, but not necessarily in an academic context. Therefore, material could be added to a task of an academic nature, to represent novel information which could to be utilised while making a decision and constructing a rationale.

This study also employs three less utilised methods of analysis in the context of explanation and argument, Rhetorical Structure Theory and two automated text parsers, HILDA and PDTB. A comparison of the differences in the type and frequency of rhetorical relations produced between experimental groups has not been carried out extensively in previous research using these methods and this will build upon the findings from the first study.

The previous research by Chi et al (1989a), Lin et al (1994) and Rosé et al (2003) suggest that learning gains are related to an increase in explanation length and the use of justificatory approaches. On this basis it will be predicted that those in the other directed group, may be prompted to write longer rationales (suggested by the findings in the first study, section 5.3.2.1), and will therefore use more argumentative structures which may enhance engagement with the material and impact retention.

The rationales will be constructed in a free text, typed format. This is to allow the focus of the task to remain on the construction of the rationale, as opposed to capturing the

rationales using an argument support tool with a graphical representation, as this may exert an additional cognitive load which could impact rationale quality and structure. The use of an argument tool may also impact upon the interaction with the learning material and therefore effect recall performance independently of experimental group.

The rhetorical and structural analyses will use the same categorisation procedure for the relations utilised in the first investigation – those of Argue, State and Analyse. As no significant differences were found in the analysis of the State and Analyse categories in the previous investigation, these relations will be excluded from hypothesis testing.

The statistical analyses relevant to each hypothesis will be reported in the findings section (6.3), followed by an extended discussion of the results in section 6.4.

6.1.1 Research Questions

Following the findings from the initial study, four research questions were generated, intended to reinforce the conclusions and to further extend the scope of investigation:

1. Does prompting for future use for a written rationale impact the perceived level of confidence in a decision based on the rationale and the perception held of the persuasive power of the rationale?

2. Does the perception of direction and future use held by an author when constructing a rationale style argument influence engagement with task material and thus recall of new information?

3. Does prompting for future use and sharing of a written rationale impact the structure and therefore quality of a rationale produced (measured by the use of Argue relations and a Toulmin based quality scheme)?

4. Does the length and structure of the rationale style argument have any bearing on perceived confidence in a decision based on the rationale?

6.1.2 Hypotheses

The following hypotheses were constructed to investigate the research questions:

H1: The prompt for future use and sharing of a written rationale will result in a higher level of perceived confidence in a decision based on that rationale compared to those who have been informed that the rationales will not be used or shared.

H2: The prompt for future use and sharing of a written rationale will result in higher ratings for perceived persuasiveness of the rationales compared to those who have been informed that the rationales will not be used or shared.

H3: The prompt for future use and sharing of a written rationale will result in greater recall of new task information compared to those who were informed that the rationales will not be used or shared.

H4: The prompt for future use and sharing of a written rationale will result in rationales that contain more Rebuttals and Counter claims (in terms of Toulmin argument elements) and thus will be more balanced compared to those who were informed that the rationales will not be used or shared.

H5: The prompt for future use and sharing of a written rationale will result in rationales that are rated as a significantly higher quality compared to those who were informed that the rationales will not be used or shared.

H6: The prompt for future use and sharing of a written rationale will result in rationales that are more complex in terms of a greater presence of 'Argue' type relations. The Argue relations that will be compared specifically include; Contrast (present in the Classical RST, PDTB and HILDA analysis frameworks) Concession (present in the PDTB and Classical RST framework) and Comparison (present in the HILDA framework).

H7: The length of the rationale and the number of Argue type relations used will have a positive relationship with the level of perceived confidence in a decision based on the rationale.

6.2 Method

6.2.1 Participants

A total of 49 participants comprised the 'other directed' (OD) group (Median age = 22 (M=26.6), with 22 females and 27 males. The second group referred to as 'self-directed' (SD) comprised of 50 participants, (Median age = 21 (M=21.7), with 25 females and 25 males. All participants were undergraduate students from the University of Bath. Allocation was randomised to each group and participant selection based on availability in response to a mail-shot.

6.2.2 Design

6.2.2.1 Independent variables:

A between subjects design was used with the direction of prompt as the independent variable with two levels:

1. **Self directed prompt** – prior to constructing the rationales participants will be informed that their rationales will be kept private.

2. **Other directed prompt** – prior to constructing the rationales participants will be informed that their rationales will be made available to others to assist in their decision making.

6.2.2.2 Dependent Variables:

The full list of dependent variables can be seen in Table 18. A full explanation of the structural analysis methods for these variables can be found in section 6.2.3.5.

	Dependent Variable	Specific Measurement
1	Decision Confidence	Likert rating scale of 1-7
2	Perceived Persuasiveness of Rationale	Likert rating scale of 1-7
3	Recall of Task Information	Gain score (post test score – pre test score)
4	Quality of Rationale	Five level Quality Framework
5	Rationale Word Length	Mean rationale length
6	Toulmin Elements	Number of Rebuttals and Counter Claims
7	Argue Type Relations	Contrast (Classical RST and PDTB, HILDA) Concession (HILDA & RST) Comparison (HILDA) relations.

Table 18 Summary of dependent variables: attitude and structure and quality.

The level of attendance to the directional prompt will also be assessed using a 7 point Likert scale; this measure is discussed in section 6.2.3.3.

6.2.3 Procedure

6.2.3.1 Pilot of Task

A pilot study of the new open ended task was conducted to explore the use of a rationale style prompt in a context intended to be similar to a psychological debate. The pilot study was run to ensure the task brief and prompts would elicit richer rationales than the QOC framework that was used to prompt participants in the first study. The debate topic was chosen from the controversial area of the nature versus nurture debate, specifically the origin of human aggression. This topic of human aggression was chosen because of its relative accessibility, yet it has rich biological and psychological research material available for both sides of the argument. This topic and the debate surrounding it, is notoriously contentious and ambiguous. It lends itself to the nature of a rationale's purpose, which requires more than a simple statement of belief or knowledge, but the reasoning behind it. This topic should be familiar to most people from a layperson perspective due to media reports and crime statistics. However, the research and scientific evidence should be less obvious to those who have not studied the topic at an academic level.

Five participants were presented with a brief containing a table of information items, arranged in two columns, one side containing evidence for a nature perspective on human aggression and one for a nurture perspective. The information in the table provided individual snippets of supporting research and formed the 'reference material' for the decision. All participants were asked to decide whether they felt humans were innately violent, and to provide a written rationale for their decision; based on the reference material where applicable. The rationales produced by the participants were balanced and provided on average around 50-80 words, compared to the first study in which the rationales averaged between 35 and 44 words. This was considered a reasonable amount on which a comprehensive content analysis could be conducted. None of the participants in the pilot test had prior knowledge of the topic at an academic level. From this pilot task there was a reasonable confidence that this decision would prompt rich rationales and result in sufficient engagement for participants regardless of prior knowledge. The rationale task was then used to conduct a larger study with additional decision evaluation measures and learning performance assessment.

6.2.3.2 Task Structure

The task took the basic structure of text based online form divided into two conditions, differing only in the inclusion of either the 'self' or the 'for others' directional prompt. Those in the SD prompt group were told that their rationales would not be shared with others (kept private) prior to making their decision and those in the OD prompt group were told that their rationale would be shared with others and used to assist in future decision making. Again, the same reasoning for suggesting future use from the first study is employed here, that the additional suggestion of future use ensures that the directional prompt may be more embedded and attended to.

The reason for eliciting a perception of future use for the rationales, rather than just seeing what people do naturally is to ensure that they do in fact hold a belief or intention that has been internalised. Otherwise the rating given for whether the rationale was self or other directed could be a factor of responding in a desired way, but would not actually have a tangible impact on their choice of reasoning styles.

Each participant was asked to complete an online form (based in Google Docs), an example of an extract from the prior knowledge capture section the rationale entry box can be seen in Figure 14.

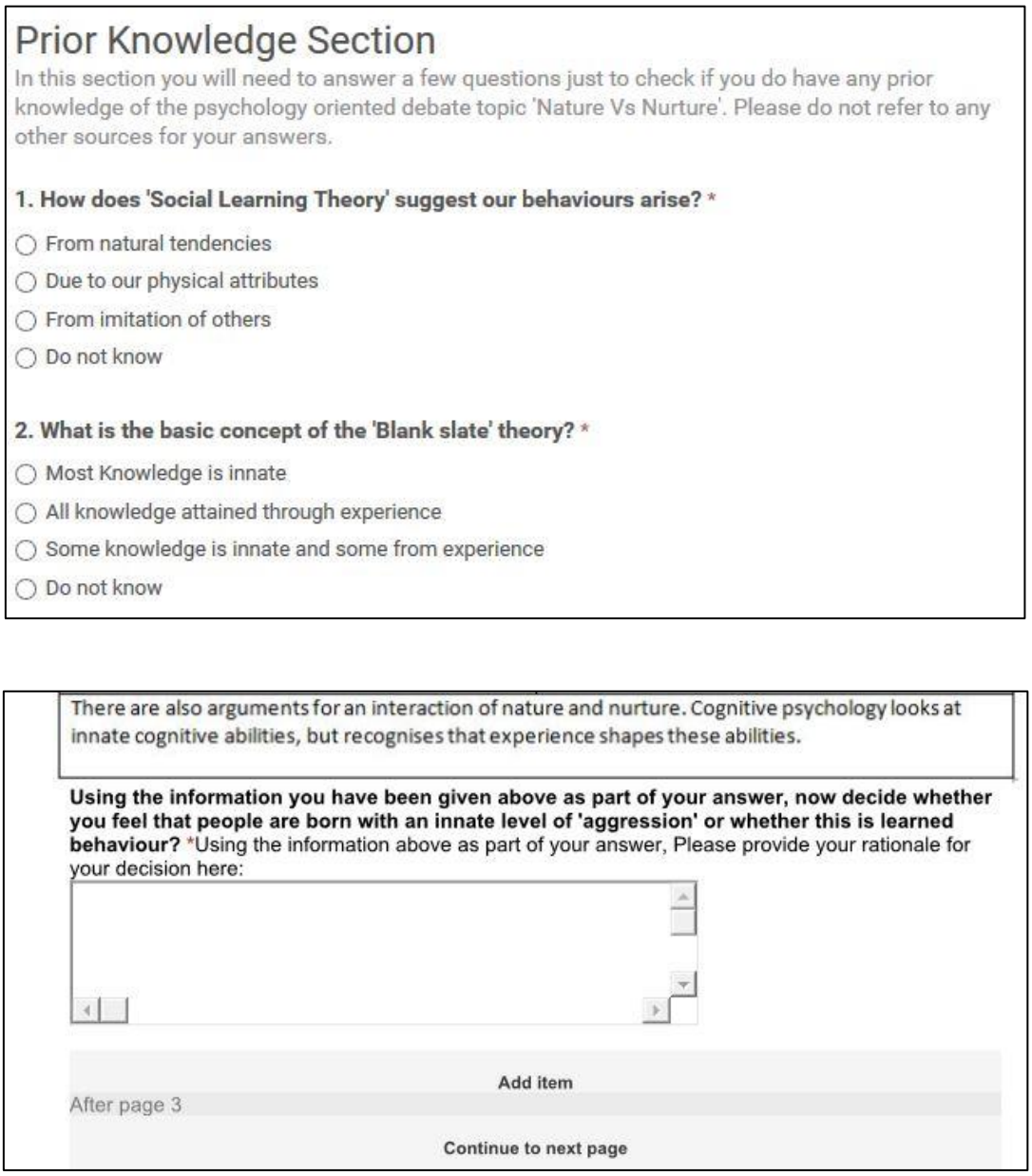

Figure 14 Screenshots of GoogleDoc (form) based decision task environment, for prior knowledge capture section (top) and rationale entry box (bottom) .

The task took approximately 15 minutes to complete, during which they were connected to the experimenter via Skype, with no video feed from the experimenter end. This was to observe that the participants completed the task without any distractions or help from others. Using groups where the perceived interaction level can be varied but the actual interaction type kept constant, helps to enable a clearer comparison of the impact of perceived direction. Feedback for both groups was non-existent while completing the task. This ensured that the impact of the expectation of future interaction was the only difference

between the context of both groups. Therefore any subsequent differences in structure or content of the produced rationales can be more reliably attributed to this intervention.

The task initially gathered demographic data including age, gender and experience with studying psychology. The next stage of the task required participants to complete a knowledge item pre-test consisting of 10 multiple choice questions. Following this, the decision task was presented. In this decision stage of the task, both groups were presented with a table containing information to support both sides of the 'nature versus nurture' debate in relation to human aggression (see Appendix 6 for full task brief). All participants were asked to decide whether they believed humans were innately violent or not, and use the information provided in the table to construct a rationale to support their decision. The resources that the participants drew from when constructing the rationales were kept limited to the brief provided and any relevant prior knowledge held by the participant. This again, was to ensure that any differences found in the construction between the groups could be more reliably attributed to the prompting of self or other directed approaches but the use of prior knowledge is not excluded.

Participants were asked to confirm their choice directly after constructing a rationale to ensure that the rationale construction was the crucial aspect of the task, prior to settling on a decision. If they were prompted to decide prior to constructing a rationale, it may inhibit the externalisation process, as participants would view the most important aspect of the task as being the decision itself and not the production of a rationale and thus would dedicate less effort to it.

Once the rationales have been completed and the decision confirmed, the participants were asked to evaluate their decision and rationales using a set of Likert based measures to gauge their attitudes. The content of these measures is discussed in the next section. In the final part of the task participants were presented with the post-test multiple choice questionnaire, consisting of reordered questions from the pre-test, to assess if any information (in addition to existing prior knowledge) could be recalled from the brief.

6.2.3.3 Decision Evaluation Measures

To address the first and second hypotheses, immediately after constructing their rationales, the participants were asked a series of questions to evaluate their decision and rationale. The post decision evaluation questions are listed below in Table 19. Responses were taken

in the form of a 7-point Likert scale. All decision evaluation measures were reversed for 50% of the sample to minimise the possibility of the data becoming skewed.

	Response Scale
1. How confident are you that you have made the best decision?	Not Very Confident at all (1) – Very Confident (7)
2. Do you think your rationale would persuade someone else to agree with your decision?	No, definitely would not (1) – Yes, definitely would (7)
3. Did you construct your rationale for yourself (to clarify your argument) or with an aim of helping others to understand your view?	For myself (1) – For Others (7)

Table 19 Decision evaluation questions.

To address the first hypothesis, participants were asked to rate on a seven point Likert scale how confident they felt in their decision. This was taken to ascertain in part, their conviction in their choice and their perception of how well they believed they had approached the task. A high score indicated a very high level of confidence that they had made the best decision, and a low score denoted a very low level of confidence was held.

Secondly, to test the next attitude based hypothesis, as an extension of the confidence measure, participants were asked to rate whether they thought their rationale would be persuasive to another person. A low score indicated that it would not be at all persuasive, and a high score indicated they thought it definitely would persuade someone else to agree with them. As persuasion is an important aspect of argumentation and with rationales being a type of free standing argument, the writers' own perception of how persuasive their arguments are, is of import here. Chittleborough and Newman (1993) propose that argumentation is a form of 'persuasive activity', when there is "an intention to either establish a proposition or persuade one or more people to accept a proposition (where such an acceptance would involve a change in belief, strength of belief or a change in behaviour)" (p. 202). They conclude that all arguments contain 'supportive' material designed to manipulative the beliefs of others and increase the likelihood of persuasion. On this basis, it could be expected that the rationales that are intentionally less self directed may contain more of these 'supportive' and persuasive elements than rationales that are more self-

directed. This measure is intended to capture how plausible the authors perceive their own arguments to be and is not a measure of how persuasive the arguments actually are. The author's intent to persuade will invariably utilise many argumentative constructs to achieve this goal and these may impact the author's attitude towards their own argument, however the actual persuasive impact on a receiver is outside the scope of this investigation.

Finally, participants were also asked to indicate on a 1-7 scale whether they had written their rationales purely as directed to themselves to clarify their view (self directed) or for others with the aim of helping them understand their view (other directed). A low score on the scale indicated a self-directed rationale was constructed and a high score indicated an other directed approach. This measure was taken primarily as an indicator of whether the prompts in each condition were successful, with the expectation that those in the OD group would rate their rationales more often as other directed than those in the SD group.

6.2.3.4 Pre Test and Post Test

To address the third hypothesis, a test of information recall was conducted with both groups to assess information retention. The pre-test and post-tests were developed to test short term recall of new information, with the pre-test forming a baseline of knowledge. Due to time restrictions and availability of participants, it was deemed appropriate to measure the short term retention of task information as an indicator of encoding. Encoding and the retention of new information are a small part of the learning puzzle and are straightforward to assess, rather than deeper procedural skill acquisition or knowledge comprehension over an extended period.

Both tests consisted of 10 multiple choice items, based directly on information present in the task brief. The information in the brief was acquired from research into the most popular sources of scientific and psychological evidence to support both sides of the nature/ nurture debate. The items are specific enough to ensure that it would be unlikely that participants would be familiar with all of the points unless a formal academic study of the topic had been undertaken. This way, extensive prior knowledge could be controlled to a certain extent, by eliminating any participants that indicated that they had formally studied this topic in an academic context. However, general knowledge of the topic gained outside of an academic course was not controlled. The extent to which participants do use prior knowledge in the task is examined in section 6.3.10.1. The items in the pre and post tests were identical, except for the order presented, which was randomised. Each test item

consisted of a question with three possible response options, and an option for 'Do not know'. One of the options represented a correct answer; the other options were constructed to appear as plausible alternatives. The correct answer was not obvious through logic, elimination or sense making processes alone. The full list of questions and multiple choice options used in the pre and post-test can be found in Appendix 7. Both the pre and post tests were given to the pilot study participants, to check no ceiling effect would occur as a result of the rationale construction. This was confirmed and some degree of difficulty still remained in the post test, with a reasonable variation between the participants. The test was designed to test basic retention of factual information and was considered sufficiently challenging to demonstrate a sensitive recall task. The test scores may help to indicate any differences in recall across groups, possibly as a result of deeper processing strategies during the task.

The raw gain scores (post score – pre score) were compared between groups and a normalised gain score calculated. The normalised gain score (g) was developed by Crouch and Mazur (2001) and was considered to be an indicator of the 'effectiveness' of a learning intervention. The formula; (g) = ((Score post) – (Score pre)) / (100%-(Score pre)), uses the percentage of correct answers for post-test and pre-test scores and accounts for pre score tests in the 'effectiveness' score (g). The use of the normalised gain in examining pre and post test scores is considered more reliable than using raw gain scores. This is due to differences in raw gain scores possibly being misleading as they represent the level of difficulty of the test items from the perspective of each individual. The gain scores do not represent actual ability or understanding, but for the purposes of this study can reasonably indicate short term memory recall of information that has been recently processed.

6.2.3.5 Content and Structure Analysis Frameworks

To address the fourth, fifth and sixth hypotheses the written rationales were analysed to reveal the structure and content using several frameworks and techniques, namely RST (a Classical approach and two automated approaches), the Toulmin model and a Toulmin based quality scheme. The frameworks and methodology for the content and structural analysis are outlined below.

6.2.3.5.1 Toulmin analysis

The fourth hypothesis states that rationales in the OD group would contain more instances of Rebuttals and Counter Claims than those in the SD group. To address this, rationales were analysed by a single analyst using the Toulmin model of argumentation. The elements within each rationale were identified and labelled, using the modified list of components below;

1. Claim: the position or claim being argued for.
2. Backing: support, justification and evidence.
3. Rebuttal/Reservation: exceptions to the claim
4. Counter-arguments: opposing claims, alternative views.

As warrants are often implicit within the text and the rationales were elicited in response to a question, these were not explicitly labelled. Similarly, the 'grounds' component of the Toulmin model was considered as semantically similar to the Backing and Claim elements and therefore omitted from the overall analysis framework. The Qualification elements (known as Modal Qualifiers) pertain to statements which indicate the strength of a claim, such as 'mostly' and 'highly' were also not considered in this analysis.

The distinction between Rebuttals and Counter claims is fundamentally that a Counter Claim is simply an alternative view, that it does not offer a description of a flaw in the original argument or a contrasting idea, just an oppositional one. Rebuttals on the other hand, are arguments that specifically address flaws in a presented argument. The Toulmin argument elements were identified in the rationales and the number of the element within each rationale recorded.

To validate the findings and ascertain the accuracy of the human analysis using the Toulmin model framework, a second independent rater used the framework to analyse the quality of a sample of 10 rationales from the corpus. The agreement levels are reported in the findings in section 6.3.6.1.

6.2.3.5.2 Quality Scoring Scale

The rationales were assessed by a single analyst to determine the quality of the argument as a whole in order to address the fifth hypothesis. The quality of each rationale was

determined using the Toulmin based quality scheme developed by Osborne et al. (2004). The quality scheme (reproduced for ease of reference in Table 20) evaluates arguments based upon the frequency of the individual Toulmin elements. The scale of the quality scheme increases from 1-5, with each increment in score indicating more structurally complex arguments, such as those arguments with more defined rebuttals and counter claims.

The theory behind the ordering of the scale is largely based on the premise that text arguments with rebuttals are, of 'better quality' than those without, because these have the ability to influence the view of the reader and indicate a higher level competency with argumentation. Kuhn (1991) argued that the ability to use rebuttals is ''the most complex skill,'' as an individual must ''integrate an original and alternative theory, arguing that the original theory is more correct'' (p. 145). The validity of the scheme and theoretical background is discussed in the literature review section 3.2.3.4.

	Description
Level 1	Consists of arguments that are a simple claim versus a counterclaim or a claim versus claim.
Level 2	Has arguments consisting of claims with data, warrants, or backing, but do not contain any rebuttals.
Level 3	Has arguments with a series of claims or counterclaims with either data, warrants, or backing with the occasional weak rebuttal.
Level 4	Shows arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counterclaims as well, but this is not necessary.
Level 5	Displays an extended argument with more than one rebuttal.

Table 20 Quality scoring scale (from Osborne et al, 2004)

Again, to ascertain the accuracy of the human analysis using the Toulmin based quality scheme, a second independent rater used the scheme to analyse the quality of a sample of 10 rationales from the corpus. The agreement levels are reported in the findings in section 6.3.7.1.

6.2.3.5.3 Rhetorical Structure Analyses

To address the sixth hypothesis the relations are again grouped in the three categories of 'state', 'analyse' and 'argue' using the method from the first investigation. This was done for all three linguistic analyses (Classical RST, PDTB Parser and HILDA Parser). Again, to account for the relations that are not included in the three categories, a fourth category of 'additional' relations will be added. The relations in this category and those of Analyse are excluded from the statistical analyses.

As previously discussed, the RST framework helps in identifying signalled and un-signalled relations. These are pieces of text that have an intended purpose that is not necessarily obvious from examining the specific word choice/discourse markers alone. The aim of using RST in partnership with other analysis methods is to begin to inform a model of argument for self-directed and other directed arguments, as well an as improved framework for rationale style argument quality analysis. RST is flexible and not restrictive in how the analyses can be used and interpreted, therefore it is a useful tool in this type of exploratory study to uncover patterns of relations across the groups. The original relations proposed by Mann and Thompson (1989) were used in the analysis and applied to all rationales. The presence and number of each relation within the rationales were recorded by a single analyst, with a smaller sample of the rationales also annotated by an independent analyst to determine inter rater reliability. The agreement levels are reported in the findings in section 6.3.8.2.

The categories of Classical RST relations are presented in Table 21 and the hypotheses testing will focus on the comparison of the use of Argue category relations between the groups. The categorisation of the Classical RST relations is adapted from the original research by Mentis et al (2009), and the previous investigation. Most notably, the Contrast relation is now included in the Argue category. This is primarily to ensure a more consistent analysis procedure as this relation is also present in the PDTB and HILDA parsers and is used to represent more argumentative text in the parser definitions (See Appendices 3 and 4). The consistencies between the parsers and the RST analysis will be discussed in chapter 10. To enable more specific and reliable hypothesis testing, the variables of Contrast and Concession are the focus of the analysis. This is based largely on the findings from the first study, which enabled the exclusion of variables which did not systematically differ between groups.

Relation Category	Classical RST Relation
'State' Relations	Background
	Elaboration
	Conjunction
	Summary
	Restatement
	Justify
'Analyse' Relations	Interpretation
	Evaluation
'Argue' Relations	Concession
	Evidence
	Antithesis
	Contrast

Table 21 Full categories of Classical RST relations

The analysis will inevitably detect relations that do not feature in these predefined categories. These relations will be labelled as 'Additional' and the findings from these will be reported in Appendix 13, and they will be used as part of a reanalysis of the approaches in Chapter 10.

6.2.3.5.4 HILDA Parser Analysis

In conjunction with a Classical RST based analysis the rationales were also analysed using both the HILDA and PDTB automated web-based parsers.

The interpretation of the HILDA parser output focussed on recording the label and number of the relation identified in each rationale. The determination of nuclearity for the EDUs was deemed to be of lesser importance to the overall analysis than the identification of the specific relation between the EDUs. Therefore, the exact nucleus and satellite labelling was omitted from the analysis. This also enables more straightforward comparisons between the analysis approaches that will be detailed in forthcoming chapters. A full discussion of the HILDA parser can be found in the literature review in section 3.1.4.

The categories of HILDA parser relations are presented in Table 22 and the hypotheses testing will focus on the use of the relevant categories. The categories were determined based on the similarity of the definitions and purpose of the relations. The relations were

also grouped with a view to categorising semantically similar or equivalent relations across all three approaches to ensure the results are comparable. The Explanation relation is described in the HILDA definitions (Appendix 3) as performing a function comparable to the Evidence relation in Classical RST. Therefore this relation was categorised as an Argue relation. To enable more specific and reliable hypothesis testing, the variables of Contrast and Comparison will be the focus of the analysis.

Relation Category	HILDA Relations
'State' Relations	Elaboration
	Background
	Attribution
	Summary
'Analyse' Relations	Cause
'Argue' Relations	Comparison
	Contrast
	Explanation

Table 22 Full categories of HILDA Parser relations

Again, any relations that are detected by the parser and are not featured in the categories will be reported in Appendix 13 and the data also examined in Chapter 10.

6.2.3.5.5 PDTB Parser Analysis

The PDTB analysis followed a similar process to the HILDA parser, with the relation label and number of occurrences being the main focus. As discussed previously, (see section 3.1.5) the PDTB parser includes the labelling of explicit and implicitly signalled relations and as this is not a feature of specific interest to the investigation; the relations are considered equal for the purposes of further analysis. Moreover, the majority of relations in the corpus were indicated as explicit. A full discussion of the PDTB parser can be found

A summary of PDTB parser relation groups can be seen in Table 23. Again, the categories were determined based on the semantic and functional similarity of the definitions. The categorisation procedure for the relation labels is discussed in the previous chapter in section 5.2.3.4.2). To enable more specific and reliable hypothesis testing, the variables of Contrast and Concession will be the focus of the analysis. This based on the findings from

the first study which enabled the exclusion of variables which did not systematically differ between groups.

Relation Category	PDTB Relations
'State' Relations	Conjunction
	Restatement
	Instantiation
'Analyse' Relations	Asynchronous
	Synchronous
	Cause
'Argue' Relations	Concession
	Alternative
	Contrast

Table 23 Full categories of PDTB Parser relations

Findings for relations not featured in these categories can be seen in Appendix 13, Table 64. The parsers may of course differ in the findings between themselves as they rely on their own classes of linguistic markers. Similarly, the parser findings may also differ in some respects to the findings using the Classical RST approach as the Classical approach relies upon judgments of plausibility made by a human analyst.

6.2.4 Rationale Length and Confidence

The final hypothesis states that the length of the rationale and the number of Argue type relations used will have a positive relationship with the level of perceived confidence. This analysis was conducted to support the findings in the previous study that indicated a link between decision confidence and rationale length. The hypothesis testing in this investigation aims to extend those findings to examine whether confidence may be related more specifically to the extended use of 'Argue' type relations.

6.3 Findings

6.3.1 Sample cleaning

The participant collection was capped at 55 for each group. Prior to analyses the raw data was cleaned to remove participants who had not written a valid rationale (having entered meaningless or irrelevant content), had stated they had academic prior knowledge of the topic or had not fully completed all of the task components. A total of five participants were removed from the SD group, three of whom had not fully completed the post test section. A fourth participant had indicated prior academic study in Psychology and the fifth had written an insufficient argument of 'yes/no' in the rationale section. Six participants were also removed from the OD group. Two of these had written insufficient arguments having recorded 'see above' and 'no time' in the rationale section. A further two participants indicated they had studied Psychology at degree level and a final participant had not completed the post test section.

6.3.2 Hypotheses 1 and 2: Decision Evaluation Measures Between Groups

The decision evaluation responses for each group were compared using a series of Mann Whitney U tests, appropriate for ordinal data. Mann Whitney analyses are often used when comparing ordinal Likert scale responses and to examine variables which have failed the Levene homogeneity of variance test. The means for both groups are summarised in Figure 15.

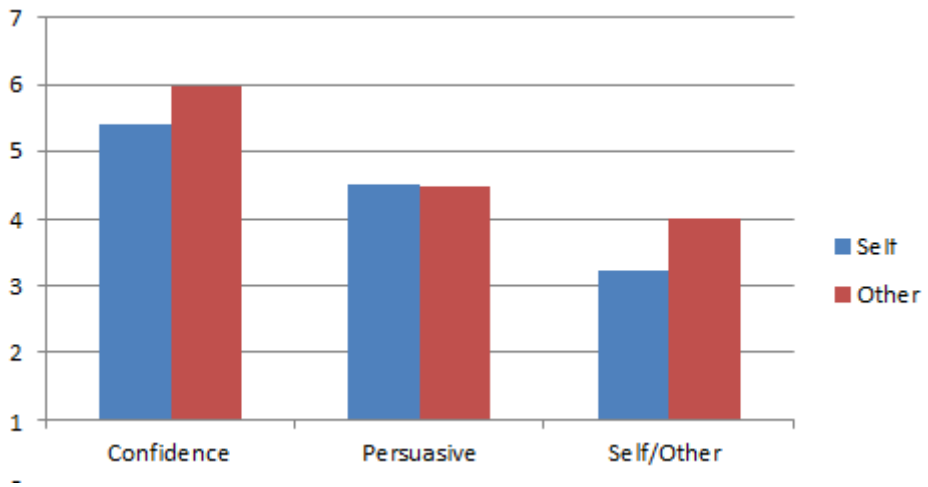

Figure 15 Comparison of Mean ratings (y-axis) for Decision Evaluation measures and Quality Scores (x-axis) between groups.

	Self Directed				Other Directed			
Measure	**Mean**	**SD**	**Mode**	**Median**	**Mean**	**SD**	**Mode**	**Median**
Confidence (1=Low)	5.42	1.20	6	6	5.98	0.97	7	6
Persuasive(1=would not persuade others)	4.52	1.59	5	5	4.49	1.39	5	5
Self/Other (1=for myself)	3.22	1.57	4	3	4.02	1.59	4	4

Table 24 Descriptive statistics for the decision evaluation measures for the SD and OD groups.

As can be clearly seen in Table 24 and Figure 15, the OD group ratings for confidence in their decision were significantly higher ('more confident') than the SD group (U = 908, p = .020, Z = -2.320, r = .25). This finding appears to support the first hypothesis. However, the second hypothesis was not supported as no significant differences between the groups were observed between the ratings for how persuasive participants thought that the rationales were (U = 1144, p = .563, Z = -.579).

6.3.3 Directional Prompt Ratings

The participants were asked whether they had constructed their rationales with others in mind, a low score on the scale indicated a self-directed rationale and a high score indicated an other directed approach. The OD group displayed a significant tendency to rate their rationales as less self directed in comparison to the SD group (U = 888, p =.016, Z = -2.402). The mode for the OD group was 7 however, the mean rating for this measure was at the midpoint of the scale, suggesting this group had a 'less self directed' approach but not necessary an extreme other directed approach. This indicates that the prompts to write rationales as either self or other directed were reasonably successful and attended to. In terms of determining the effectiveness of the OD and SD prompts the Likert ratings for this measure were examined. A total of 52% of those in the SD group rated their rationales as more self-directed (score of 1-3) and only 14% rated as other directed (score of 5-7). In comparison, in the OD group 43% rated their rationales as more self-directed but 39% of those in the OD group rated their rationales as more other directed. This indicates that the

prompts were largely successful in facilitating a differing perception of the purpose of the rationales, and this therefore may be a factor in the observed structural differences.

A post-hoc comparison of the ratings given for direction within the two groups was also conducted to determine the effectiveness of the prompts. The ratings at the midpoint of the scale were removed for both groups and the rationales and the ratings for more self directed and more other directed were compared within groups. The ratings and distributions of participants can be seen in Table 25.

	N
SD Group Rating 1-3	26
SD Group Rating 5-7	8
OD Group Rating 1-3	21
OD Group Rating 5-7	19

Table 25 Directional group distribution for the SD and OD groups.

The results did not reveal any systematic differences in argumentative structures present or attitude towards the decisions based on these extremes. This finding perhaps indicates that those who claimed to construct 'other directed' rationales in the SD group did not in fact adopt the same strategies as those in the OD group overall. This may suggest that the outward expression of attitude and the rating that a rationale is other directed may not be a straightforward reflection of internal strategies. It is in fact, the elements of the context such as the presence of a prompt to suggest future use or direction that changes the strategies adopted and this alters the externalised rationale. The addition of a prompt may help to induce a perception of a future use and direction for the externalised rationale and this changes the approach from merely thinking you have written in a less self directed manner to adopting strategies that are more akin to actually doing so.

6.3.4 Hypothesis 3: Task Information Recall Between Groups

An analysis of task performance, specifically in terms of recall of information items from the task brief was conducted to ascertain if an enhanced learning effect may be present for those in the self or other directed rationale groups. Pre and Post test scores were calculated and a raw gain score produced (Post-test – Pre-test). In addition, the normalised gain score

was calculated and added to the analysis. The descriptive data for the pre and post scores for both groups, and the raw gain scores are summarised in Table 26.

	Group	Mean	*SD*
Pre-score (max 10)	SD	6.00	2.25
	OD	5.14	2.49
Post-score (max 10)	SD	7.54	2.36
	OD	7.57	1.62
Raw Gain (post-pre)	SD	1.54	1.97
	OD	2.43	2.34
Normalised Gain score	SD	0.36	0.45
	OD	0.44	0.29

Table 26 Mean pre and post test and recall performance scores between groups.

A one sample t-test was conducted to determine whether the task intervention itself produced significant differences between pre and post test scores within each group. The OD group demonstrated a significant difference between pre (M= 5.14, SD=2.49) and post-test (M=7.57, SD=1.6) scores ($t(48)=-7.276$, $p<.01$). The SD group also demonstrated a significant difference between pre (M=6, SD=2.25) and post-test (M=7.54, SD=2.36) scores ($t(49)=-5.524$, $p<.01$).

The modest post-score mean of approximately 7.5 for both groups indicated a ceiling effect was not present. There was no significant difference found between the pre-scores in the SD and OD groups. This is reassuring, indicating a comparable prior knowledge level for both groups and therefore a higher level of confidence in attributing any differences in performance to the differing experimental conditions.

To test the third hypothesis, a t-test analysis was conducted between the groups and revealed that the raw gain scores (post score – pre score) were significantly different ($F= .548$, $t(97)= 2.047$ $p= .043$, $d = .41$), with the OD group having the greatest raw gain. This finding supports the hypothesis, however, there was no significant difference between the groups for the normalised gain scores ($F= 11.714$, ($t(97) = 1.017$, $p = .312$).

6.3.5 Quality and Structural Analysis

A sample of rationales, from both the OD and SD groups, can be found in Appendix 9.

6.3.5.1 Rationale Length

The descriptive statistics for rationale length in both groups are summarised in Table 27. The word count data failed the homogeneity of variance test and therefore a Mann Whitney analysis was performed on the mean scores.

	Rationale Length		
Group	**Mean**	***SD***	**Median**
SD	66.56	36.95	2
OD	86.86	51.57	3

Table 27 Mean rationale length for the SD and OD groups.

Table 27 shows a difference in the average rationale word count for both groups, however this was not statistically significant (U = 967.5, p = .071, Z = -1.802). This would suggest that any subsequent differences found in rationale structures between the groups would be more reliably attributed to there being a greater proportion of the structures in either group, rather than it simply being a result of more words written.

6.3.6 Hypothesis 4: Toulmin Element Comparison Between Groups

The Toulmin analysis generated four variables which were identified within the rationales according to the frequency of the elements present in the text. The variables and means for the components in each group are summarised in Figure 16 and Table 28. The medians for each relation are reported for completeness and where non-parametric analyses are used. However, due to the high number of zero values in the data the medians are less informative than the means. Therefore, for visual representation of the data, the mean value will be used.

Toulmin Element	Group	Mean	*SD*	Median
Claims	SD	1.86	0.9	2
	OD	1.69	0.96	1
Counterclaims	SD	0.28	0.45	0
	OD	0.53	0.54	1
Backing	SD	2.28	1.43	2
	OD	4.49	3.37	4
Rebuttals	SD	0.26	0.49	0
	OD	0.41	0.54	0

Table 28 Descriptive statistics for Toulmin analysis of the SD and OD groups

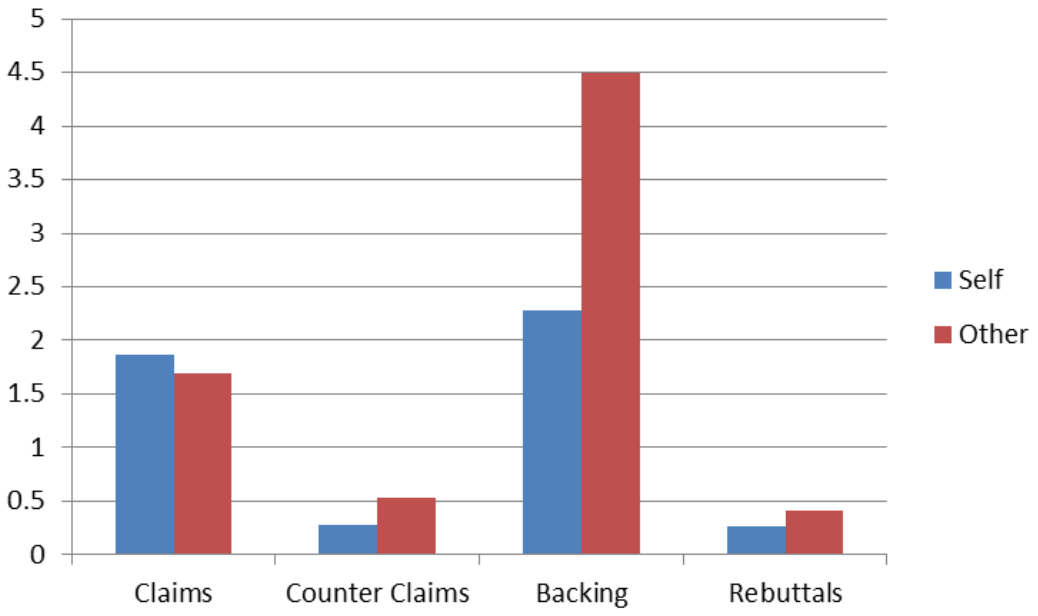

Figure 16 Comparison of Means (y-axis) between groups for individual Toulmin elements (x-axis) within the rationales.

As can be seen from the graph, Claims with Backing make up the bulk of the arguments as expected, with the more complex forms of arguments, Counter Claims closely followed by Rebuttals as the less frequent elements. The hypothesis is partially supported by the finding

of a significant difference between the SD and OD group (U =936, p = .017, Z = -2.382, r = .24) in the use of Counter Claims, with the OD group containing more of these elements. The Counter Claims variable failed the homogeneity of variance test; therefore a Mann Whitney U test was performed. However, no significant difference was found between the groups in the use of Rebuttals (U = 1047.5, p = .124, Z = -1.540). The variables that fail the homogeneity of variance test should be considered with caution.

6.3.6.1 Rater Agreement: Toulmin Analysis

The large sample size and extensive nature of the analysis meant that it was not possible to have multiple analysts for the entire corpus. Instead, a small sample of 10 randomly selected rationales was assessed using an independent rater for comparison. This was to improve the use of the analysis approaches and ensure a more systematic approach. In order to be more confident regarding the analyses, a co-analyst for the entire corpus would be needed to ensure that the identifications were consistent throughout.

A sample of 10 rationales was randomly selected and analysed using the Toulmin model approach by a second, independent rater. The Kappa coefficient between the ratings was .721 indicating a good level of agreement between the raters.

6.3.7 Hypothesis 5: Quality Comparison Between Groups

The descriptive statistics for the Quality ratings can be seen in Table 29. The fifth hypothesis is supported as the Quality scale scores assigned to the rationales were found to significantly different between groups (F= .587, t(97)= 3.782, p=.000, d = .77), with those in the OD group having higher mean scores for quality compared to the SD group.

	Quality Scale Score		
Group	**Mean**	***SD***	***Median***
SD	2.24	.82	2
OD	2.94	1.00	3

Table 29 Descriptive statistics for Quality scale ratings in the SD and OD groups.

6.3.7.1 Rater Agreement: Quality Analysis

A sample of 10 rationales was randomly selected and evaluated using the Toulmin based quality scheme by a second, independent rater. The Kappa coefficient was .855 indicating a very good level of agreement between the raters.

6.3.8 Hypothesis 6: Argue Relations Comparison Between Groups

6.3.8.1 Classical RST Analysis

The Rationales were deconstructed into tree style diagrams using a Classical RST annotation Tool. Examples of each relation identified within the corpus are available in Appendix 10. Spelling and punctuation errors within the rationales were corrected prior to being analysed by both discourse parsers to reduce errors based on these features.

The Classical RST findings in each group were compared using a t-test analysis but where appropriate, a Mann Whitney test was performed. The descriptive statistics for the Classical RST analysis are summarised in Table 30 and visually presented in Figure 17 (the full summary for the Classical RST Relation categories is available in Appendix 13, in Table 62 and Table 63). The medians for each relation are reported for completeness and where non-parametric analyses are used. Again, due to the high number of zero values in the data the medians are less informative, therefore, the mean value will be used for visual representation of the data. In terms of hypothesis testing, the Concession and Contrast relations are the variables of interest.

	Group	***Mean***	***SD***	***Median***
Concession	SD	0.82	0.75	1
	OD	1.29	1.02	1
Contrast	SD	0.16	0.55	0
	OD	0.24	0.66	0

Table 30 Descriptive statistics for Classical RST Argue category relations

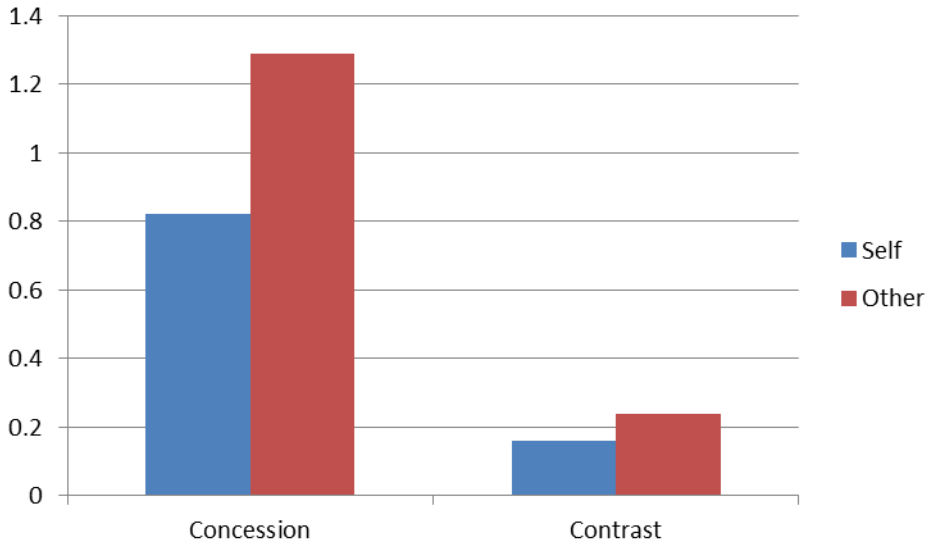

Figure 17 Means (y-axis) for Classical RST Argue category relations between groups.

The number of Concession relations between the groups appears to be the most marked difference. This was found to be statistically significant (F = 1.950, t(97) = 2.594, p=.011, d = .53) with the OD group containing a higher number of Concession relations compared to the SD group, thus supporting the hypothesis. However, no significant difference was found in the use of Contrast relations (U = 1173, p = .486, Z = -.697).

6.3.8.2 Rater Agreement: Classical RST Analysis

A sample of 10 rationales was randomly selected and analysed by a second, independent rater using the Classical RST framework. The Kappa coefficient between the ratings was .795 indicating a good level of agreement between the raters.

6.3.8.3 PDTB Parser Analysis

Examples of the relations identified in the corpus are briefly summarised in Appendix 11 to give an indication of the types of discourse markers that give rise to the assignment of a particular relation.

The automated analysis findings using the PDTB parser were also analysed using a series of t-tests, with the descriptive statistics for the relations identified in each group summarised in Table 31 and presented visually in Figure 18 (the full summary of the PDTB findings for

all categories of relation can be found in Appendix 13, Table 64). As with the findings for the Classical RST analysis, the means are used when visually representing the data.

	Group	Mean	*SD*	Median
Concession	SD	0.08	0.27	0
	OD	0.18	0.49	0
Contrast	SD	0.7	0.79	1
	OD	1.12	0.95	1

Table 31 Descriptive statistics for PDTB Parser Argue category relations.

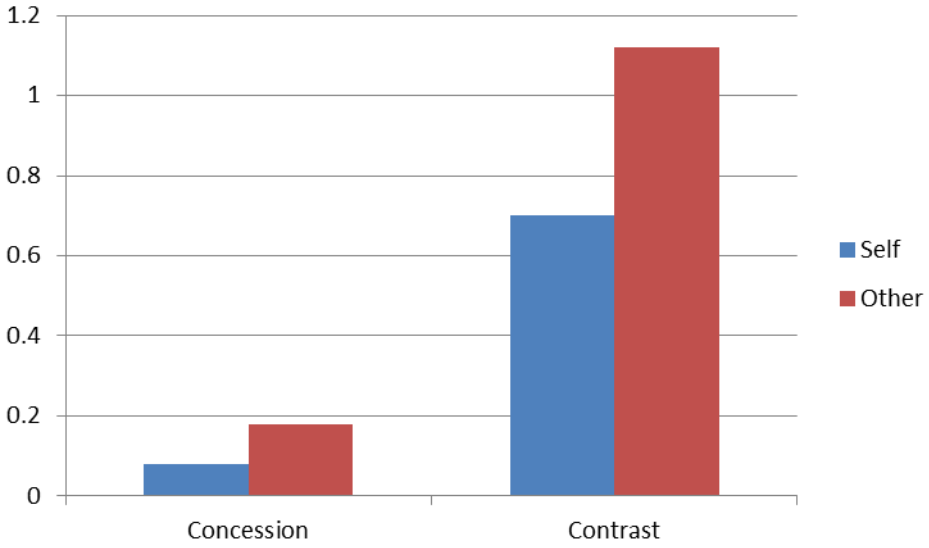

Figure 18 Means (y-axis) of PDTB parser Argue category relations (x-axis) between groups.

In terms of hypothesis testing, the Concession and Contrast relations were the Argue relations of interest, with the Alternative relations excluded from the analysis. The Concession relations were not significantly different between the groups and the hypothesis was not supported in this case. This variable failed the Levene test and was examined using a Mann Whitney U analysis (U = 1144, p = .298, Z = -1.040). However, the use of the Contrast relations was found to be significantly different between the groups (F = .111, t(97)=2.410 p=.018, d = .48), with the OD group rationales contain more of these relations than the SD group.

6.3.8.4 HILDA Parser Analysis

Examples of the relations identified in the corpus are briefly summarised in Appendix 12 to give an indication of the types of discourse markers that give rise to the assignment of a particular relation.

The output from the automated HILDA analysis was also compared between groups to examine the use of the Argue category of relations. The descriptive statistics for the Argue category of relations in each group are summarised Table 32 the means represented visually in Figure 19. Comparison and Contrasts were the variables of interest for hypothesis testing (a full summary for the HILDA relation findings for all categories of relation is available in Appendix 13, Table 65).

	Group	Mean	*SD*	Median
Comparison	SD	0.02	0.14	0
	OD	0.16	0.43	0
Contrast	SD	0.36	0.56	0
	OD	0.53	0.58	0

Table 32 Descriptive statistics for HILDA Parser Argue category relations

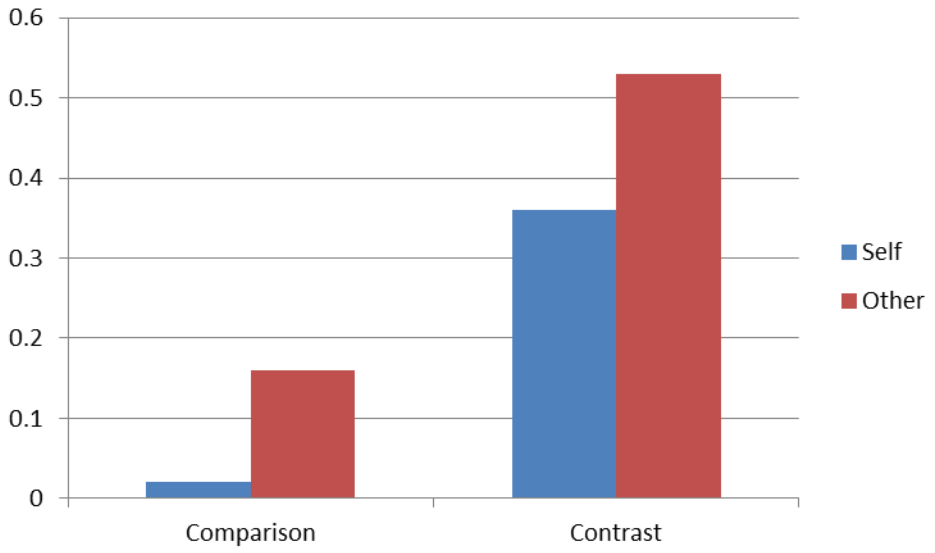

Figure 19 Means (y-axis) of HILDA parser Argue category relations (x-axis) between groups.

The mean number of Comparison and Contrast relations appears to differ extensively between groups. However, only the number of Comparison relations was found to be significantly different between the groups ($F = 24.148$, $t(97)=2.257$ $p=.026$, $d = .44$), thus partially supporting the hypothesis. This variable failed the homogeneity of variance test and this prompted a further comparison using a Mann Whitney U ($U = 1074$, $Z = -2.238$, $p = .025$, $r = .22$), which confirmed the statistical significance.

6.3.9 Hypothesis 7: Confidence and Rationale Structure

The final hypothesis states that the length of the rationale and the number of Argue type relations used will have a positive relationship with the level of perceived confidence. The confidence ratings for each group were correlated with rationale length and the Argue category of relations for all three approaches. Contrary to the findings in the first study, confidence ratings did not correlate with rationale length for either group, or when the groups were combined. In the case of Argue relations and confidence ratings, a weak positive correlation was found between confidence ratings and Classical Concession relations in the SD group only ($r_s (50) = .307$, $p<.01$).

6.3.10 Post-Hoc Comparisons

A number of post-hoc comparisons between the groups were conducted in the interest of completeness and to validate the exclusion of the State and Analyse categories from the hypotheses.

6.3.10.1 Concept Origin

Participants were instructed to only use the task information brief to provide backing for their argument, however, the use of prior knowledge was not controlled as the topic may be familiar, and prior beliefs held. To assess the extent to which information from outside the brief was used within the rationales, a further examination of the content of the rationales used a method adapted from research carried out by Coleman (1998). The 'concepts' were identified within each piece of text. The concepts are individual terms which refer to scientific or real world concepts or objects included in the argument. A concept is counted as many times as it appears, even if repeated. Synonyms of concepts that featured in the brief information were counted and considered to be from the brief. Whether or not the concept was present in the information brief was also recorded, to ascertain how much of

information in the rationales was derived from the brief and how much prior knowledge was incorporated into the argument. The descriptive statistics for the concepts and origin features are summarised in Table 33.

	Group	**Mean**	***SD***
Total Concepts used	OD	14.90	9.67
	SD	11.44	6.16
Concepts from Outside the Brief	OD	1.45	1.93
	SD	1.14	1.47

Table 33 Concept use and origin between groups.

The means for concept use and origin also appear to be relatively similar between the groups and subsequently no statistically significant differences were found. It appears that participants used information within the task brief in the majority of their arguments, which a small portion of one or two concepts that may have been existing knowledge.

6.3.10.2 Decision Evaluation: Within Group Correlations

To ascertain the relationships between the decision evaluation measures and the directional ratings a Kendall tau-b correlational analysis was performed. This test is primarily used in case of non-parametric data distributions and for ordinal measures. This was considered worthwhile to confirm that the evaluation questions were indeed measuring different attitudes.

No correlations were found between the decision evaluation measures within the OD group. Within the SD group, the confidence ratings correlated significantly with how persuasive the participants thought their rationales would be ($tb=.408$, $N=50$, $p<.01$). The ratings for how persuasive the rationales were considered to be also correlated significantly with whether the rationales were self or other directed ($tb=.377$, $N=50$, $p<.01$). These results may indicate that those in the SD group who constructed less self directed rationales considered the arguments to be more persuasive to others and also held a higher perception of their level of confidence. However, these findings are post hoc and as a higher number of comparisons were made, the findings are not significant when the accepted significance level was corrected using the Bonferroni method.

6.3.10.3 Use of State and Analyse Category Relations Between Groups.

To confirm that the exclusion of the State and Analyse categories from the main analysis was valid, and to ensure that no significant findings were overlooked, a post hoc comparison of these categories between the groups was conducted and the findings summarised below. The findings for all three approaches for these categories can be seen in Appendix 13 (Table 62, Table 63, Table 64 and Table 65).

For the Classical RST, PDTB and HILDA State categories of relations, no significant differences were found between the groups with the exception of the Classical RST Background relation ($U = 705.5$, $p = .000$, $Z = -4.924$, $r = .45$). The SD group appeared to contain significantly more of this type of relation than the OD group.

6.4 Discussion

This study expanded on the initial findings of the exploratory study which demonstrated that other directed arguments appeared to be lengthier and more complex than self-directed rationales. Additionally, this increase in rationale length appeared to coincide with an increase in confidence ratings, indicating an intriguing relationship. The decision task differed for the current study, being a less structured topic to better reflect the more typical questions that may be encountered in a real world setting. Participants were also able to draw information from any prior knowledge or beliefs if they so wished, though they were not specifically asked to do so.

A significant amount of data was generated from the research that was extensively analysed using a variety of frameworks and tools to assess argument structure and coherence. Therefore this section will focus on summarising, drawing the results together and discussing the implications. The hypotheses that underpinned the study will be discussed in this section followed by a consideration of the limitations and conclusions.

6.4.1 Hypothesis 1: Confidence Comparison Between Groups

The self and other directed groups differed significantly in the ratings given for confidence in the decision. The OD group appeared to report higher perceived confidence in their decisions. The results conform to the expectations raised from the literature and the results

support the hypothesis. The mechanisms that underlie the effects of the self-explanation effect including an increase in the depth of processing may be responsible in part for the observed differences in confidence ratings.

This notion of confidence being enhanced as a result of an increased depth of processing is a plausible explanation for the observed differences. Those in the other directed group appear to have approached the rationale construction task differently (as evidenced by significant structural differences). If these structures are evidence of different reasoning styles it is not unreasonable to infer that the production of these may alter the attitude held towards the argument. The increased frequency of argument style relations such as Contrasts, are arguably more complex and require more evaluation of the information compared to simpler Elaboration relations.

People do have a tendency to attend more to evidence that supports their belief, even if the evidence is flawed or if contradicting evidence is readily available. This tendency is known as belief perseverance (Guenther & Alicke, 2008), a robust effect in social psychology. This tendency may impact on perceived confidence in a decision, as only supporting evidence is attended to, leading to a biased perception of support for a belief. This tendency could arguably have occurred in either group and thus is probably not a substantial explanation for the effects observed. An attempt was made to minimise the chance of participants creating a totally biased argument by delaying the confirmation of their decision until after they had constructed a rationale.

The types of argumentation techniques used within the rationales may be responsible for the differences in confidence between the groups. Worthy of note is that those in the OD group indicated that the rationales were written as less self directed in comparison to the SD group, suggesting the approaches were indeed different, perhaps due to perceived differences in the rationale purpose.

6.4.2 Hypothesis 2: Persuasion Comparison Between Groups

The second hypothesis was not supported as no significant difference was found between groups for the ratings of perceived persuasiveness. The trends observed in the results (Table 24) do offer an interesting perspective on how confidence in a decision, as well as positive evaluations on the persuasiveness of the arguments generated could be further

facilitated by prompting for an other directed approach. However, the persuasiveness measure is not indicative of the actual persuasive power of the rationale, which is measured by the impact on the receiver and not generally considered as a being a function of author intention. This means that no conclusions can be drawn regarding whether the OD group rationales were actually more persuasive, only that there is a trend (though not significant) that OD authors may view their own arguments as potentially persuasive to others.

6.4.3 Hypothesis 3: Task Information Recall Between Groups

This hypothesis was partially supported by a significant difference in the raw gain (post test score – pre test score) scores. The raw gain scores were significantly higher in the OD group. However, this difference did not persist for the normalised gain scores. On this basis the hypothesis is not robustly supported, but the findings are in the expected direction. The results suggest that an enhanced encoding effect may be present in favour of those who construct other directed rationales. It may be that there is potential for enhancing recall of information by prompting for other directed explanations. As the OD group appeared to have a higher frequency of Classical Concession and Contrast (PDTB) relations, it could be considered that these more complex forms of argument result in a deeper engagement and processing of the learning material. This behaviour may facilitate the encoding of new information and thus have the potential to impact recall of information at a later date.

The issue of externalisation is also worth considering here, although constructing rationales appears to be successful as part of task, and also appears to assist short term recall of new information, it does not fully account for the internal processes which may be occurring simultaneously with rationale construction. Previous research has indicated that 'working memory dumps' and talk aloud protocols are not as effective as written explanations (Schworm & Renkl, 2007), and neither is simply planning explanations without writing them (Sieck & Yates, 1997). This would suggest that the construction and explicit organisation of arguments, which are largely coherent and often narrative in nature, are the key factors in the observed benefits.

The presence of a significant difference in raw recall gain scores between the OD and SD groups is an encouraging indication that a trend may exist, but whether or not this can be enhanced further is unclear. It may be that the presence of more complex argument structures with the OD group rationales was a result of increased interaction with the new

material. It may be ambitious to attribute structural explanations at this this stage, but it could be possible that the use of Contrasts, Concessions and Comparisons within a rationale may increase interaction with and processing of the material available and therefore be partially responsible for any differences in recall scores.

The difficultly with interpretation of the data is the absence of a learning motivation in the task procedure. Therefore it seems ambitious to attempt to infer reliable causal relationships between recall performance and specific structural features within the rationales.

6.4.4 Hypothesis 4: Toulmin Element Comparison Between Groups

As the reported ratings given for other or self-directed approaches were significantly different between groups, it was considered appropriate to compare both groups for structural differences in order to address the second hypothesis.

The manual Toulmin analysis revealed a significant difference in the number of Counter Claims between the SD and OD group. Just over half of the rationales in the OD group contained Counter Claims compared to just over a quarter in the SD group. As the Toulmin model is not intended to be descriptive of how people argue, but a description of the structures, caution must be taken into attributing a causal nature to the findings. However, the consistency of the structural differences revealed throughout all of the analyses seems to indicate a genuine trend. Interestingly, discussions on argument strength and quality by Kuhn, suggest that successful arguments consist of four components; a statement, an alternative, a rebuttal of the alternative, and a final counter argument and rebuttal. The use of rebuttals was considered as the strongest form of argument, given that it requires a balanced and evaluative view of both sides of the argument and anticipates any attempts to discredit the original claim, thus avoiding circular arguments. The discovery of a greater proportion of Counter Claims within the other directed group rationales could therefore be interpreted as an indication of greater argumentative strength and power within this group.

6.4.5 Hypothesis 5: Quality Comparison Between Groups

Firstly, it is important to note that a Mann Whitney analysis comparing the rationale lengths between groups revealed no significant difference. This is worth bearing in mind when

attributing possible causes of structural differences in the rationales, or attitudes towards the quality of the rationales. The differences found between the OD and SD groups would ideally not be a result of simply writing 'more' but of writing differently.

The use of the Toulmin based quality scheme, to manually assess the argument quality, revealed that the OD group arguments were of significantly higher quality than the SD group. The finding suggests that constructing a rationale in a perceived other directed context may result in a significantly different structure to be externalised in comparison to a self directed approach. This appears to support the hypothesis.

The higher argument quality levels indicate that these rationales contain a greater number of Rebuttals and Counter Claims. As Contrasts were also found to be significantly different between groups in conjunction with the quality scores, the assumption that the rationales with these relations represent better quality arguments is supported. The implication that Contrast relations indicate elements that have the argumentative purpose of 'rebuttal' is examined further in section 10.2.4.

6.4.6 Hypothesis 6: Argue Relations Comparison Between Groups

The sixth hypothesis predicted that the OD groups would construct rationales with more Argue category relations than the SD group. Evidence for this can be seen in the increased frequency of Classical RST Concession relations within the OD groups (82% of rationales contained a Concession), indicating that those who constructed other directed rationales appeared to favour more complex argument structures. A Concession based argument requires an evaluation of a concept and identification of possible flaws in the argument, as opposed to an Elaboration, which usually offers additional information for a concept mentioned with no further analysis or interpretation.

The results of the automated parsers concurred to some extent with the Classical analyses. There was a significant difference in the number of Comparison relations between the groups, as detected by the HILDA parser. However, the finding of a significant difference for the HILDA Comparison relation between the groups does not appear to concur with the PDTB or Classical Contrast relation findings. The definitions in the Parser frameworks demonstrate a discrepancy in the labelling that is worth noting here. As can be seen in Appendix 4, the Contrast and Concession relations are grouped together under the

Comparison class, however, the HILDA Parser labels Comparison (see Appendix 3) as a separate construct to those of Contrast and Concession. Both Contrast and Concession are linked under the Contrast class in the HILDA parser, suggesting, contrary to the original RST framework, that these two relations perform represent similar functions in the text. The frequency of Comparison relations identified is considerably lower (15% and 2%) than the Contrast relations identified in the PDTB parser (78% and 58%) so the direct mapping of the Contrast and Comparison relations may not be straightforward. For a discussion of the limitations of the parsers see section 11.7.3.

The frequency of these relations overall was very low however, so conclusions are difficult to draw from this observation alone. The most intriguing structural difference detected between the groups was the frequency of Contrast relations detected by the PDTB parser. In the Parser definition (Appendix 4), the Contrast and Concession relations are included in a single class labelled 'Comparison.' This suggests that the developers viewed these relations are performing a similar function in the text. Thus in this instance the findings using the PDTB parser could be considered to concur with the Classical RST analysis. A Contrast relation may be a considered a complex form of argument as it will involve a comparison between two claims or concepts with a possible intention of increasing a favourable perception of one of the elements. The apparent trend of more Contrast type relations existing in the OD rationales is a tentative indication perhaps of the greater persuasive nature and argumentative complexity of these rationales.

Whether the Contrast relation is in fact utilised as an intentionally persuasive argument is a concept worth further consideration, as in the original Classical RST framework the Contrast relation is listed as a neutral subject matter relation. This will undergo further examination later in the thesis.

6.4.7 Hypothesis 7: Confidence and Rationale Structure

The final hypothesis states that the length of the rationale and the number of Argue type relations used will have a positive relationship with the level of perceived confidence. The confidence ratings for each group were correlated with rationale length and the Argue category of relations for all three approaches. The results did not support the finding from the previous study that revealed a relationship between confidence and rationale length. It could be argued that the observed increase in confidence for the other directed group may

simply be a result of the 'perceived effort effect' (Sieck & Yates, 1997). If this effect were the responsible mechanism for increased confidence, the lengths of the rationales would correlate positively with confidence ratings. However, no such correlation was observed. It may be that the determinants of confidence for those in the OD group are more complex than simply relying on elaborative elements as an indicator of argument quality.

Additionally, only a single correlation was found in the SD group between the use of Concession relations in the Classical RST analysis and confidence ratings. As no other significant correlations were found it can be concluded that this hypothesis, based on the initial findings from chapter 5, is not supported.

6.4.8 Post-hoc Comparisons

A number of post-hoc comparisons between the groups were conducted in the interest of completeness and to validate the exclusion of the State and Analyse categories from the hypotheses.

Upon examining the Classical RST analysis results, the OD and SD groups were also found to differ significantly in the number of Background relations with the SD group rationales containing more instances of this relation. The use of Background relations may be a less demanding or complex approach to supporting an argument compared to using Contrasts or Rebuttals. No other differences were found between the groups in the use of State relations in the Classical RST analysis. Additionally, no State relations were found to differ using the PDTB or HILDA approaches, therefore it may well be the variation of human analysis that is responsible for this finding. Additionally, the high number of comparisons of relation types conducted between groups may invalidate this finding.

Overall, the lack of statistically significant differences for the State and Analyse category of relations confirms that the exclusion of these relations from consideration, on the basis of the infrequency of these relations in the first investigation, was valid.

A further area of interest addressed in the post-hoc analysis is how the evaluation measures related to the perception of direction for the rationale. A particularly interesting finding is the apparent tendency for those who had constructed other directed rationales to also perceive their rationales as more persuasive. This correlation was found only within the SD

group. The measure for the author's perception of persuasion is not representative of how persuasive the argument will actually be to receiver, but simply how persuasive the author thinks it may be. As no significant differences were found in the ratings for persuasiveness between the OD and SD groups nor were any significant correlations found for decision measures in the OD group, it is difficult to draw any explanatory conclusions from this finding.

6.4.9 Limitations

One of the most pertinent findings from the investigation is the lack of extreme difference in the average ratings for direction between the groups. The OD group, on average rated direction of their rationales around the midpoint of scale. This suggests that the OD group adopted a 'less self directed' approach as a whole rather than a 'fully other directed' approach. However, an examination of the proportion of ratings within the groups revealed that 43% of participants in the OD group rated their rationales as more other directed compared to only 14% in the SD group. It may be that the directional prompt used prior to rationale construction (that stated that rationales may be 'used to assist others') is too distinct in purpose from the actual direction rating question which asked participants if the rationales were constructed with the aim of 'helping others to understand your view.' The rating question would have been better equipped to fully assess the effectiveness of the prompt if the wording had mirrored the original text.

It may be that the differences between the OD and SD group in terms of structures are in fact due to a consideration that their rationales will be used by others and this may be why the ratings for direction are not as polarised as they ought to be. Similarly, the perception of future use that the SD group authors held was not measured, only the intended direction. This may be an issue as there are arguably reasons for a future use that someone may generate in a reflective, self directed argument such as for revision purposes or to record a position for future reference. The perception of future use and the possible variants needs to be clarified as it has been inextricably linked to the perception of direction in this empirical work.

There is also a chance that the perception held of writing as self or other directed, may generally default to 'other' directed when using an online environment. This is possibly as a result of prior experience of using the internet as a platform for social interaction.

Participants may view any or indeed all online activity as interactive to an extent. Therefore the use of GoogleDocs as the task delivery environment may carry with it a sense of interactivity regardless of any prompts. Participants may also rate their rationales as being more self-directed, if they considered them to be of poor quality or less convincing, so as to 'justify' the quality of the rationales. However, if this was the case many more participants would have rated their rationales as one or the other extremes. In fact, the ratings for all decision evaluation measures for both groups, were of a similar distribution and with a wide range of scores within each group. Overall, the significant differences in ratings for self or other directed rationales confirmed that the prompts were successful in changing the perceived direction of the rationales.

There are of course several reasons to interpret the data with some caution. There should always be a level of scepticism when interpreting self-report data. There may be a possibility that participants may have responded in a way that they felt would be expected, such as rating confidence as higher in response to the prompts that suggested their rationales would be shared. This may be due to a desire to appear more confident regardless of genuine feeling. However, the absence of any performance rewards or feedback will have possibly minimised these types of reactions. Similarly, it could be posited that participants thought they were expected to perform in a particular way, however both groups were given the same loose definition of a rationale and directed to the same task materials to aid in the rationale construction. Any systematic differences in the structures within the rationales could thus be more reliably attributed to differing approaches resulting from different perspectives on the direction (or purpose) of the rationales.

In terms of generalising the findings from the learning performance data there are a number of considerations. The learning component of the task is somewhat limited for the participants in terms of the time spent with the material. Almost all rationales were constructed within a 10 minute period, which is a significantly shorter 'learning' intervention than has been conducted in previous research. However, the brief contained a reasonably small amount of information and if participants were forced to spend extensive time with the brief, a ceiling effect may have occurred. For the purposes of examining whether other directed rationales may have an impact on information retention and recall, this study indicates that this is an area worth investigating further.

Additionally, some of the performance scores improved from the pre-test to the post test scores but not all, suggesting that some participants did not engage fully with task or that participants simply did not feel the task was to memorise the new information. The task did not purposefully elicit a motivation for learning; therefore any additional information recalled in the post score test would have been encoded implicitly, to some extent, as a by-product of the rationale construction process. This lack of learning motivation may have resulted in more noise in the data and less consistency, as participants were focussed on the goal of rationale construction rather than learning and therefore attention to the task brief was not monitored or controlled.

The pre-test element of the task was introduced as a prior knowledge check, with no indication that a post-test would follow. This removed the motivation to memorise any newly presented information. This minimised the chances of a ceiling effect and helped to ensure that any trend in differences with engagement or performance could be more reliably attributed to the experimental condition. To investigate the effect of self or other directed rationales on recall, a more explicitly motivated learning task would need to be devised. However, this type of demanding task and shift in focus may involve a trade off with rationale quality.

The pre-test questions at the start of the task could also be argued to have affected confidence if the participants found the questions difficult or did not know the answers. To examine this, an analysis of low pre scorers versus high scorers within groups was conducted. This did not reveal a significant difference in confidence, indicating that low pre-test scores did not measurably impact upon confidence. Confidence ratings appear to be a result of the actual decision making and rationale construction aspect of the task as intended, rather than residual feelings in response to the pre-test.

The concept of subsumptive constraints discussed in the literature review may be worth considering in this context. When forcing people to create an argument, the subsumptive constraints may have a detrimental effect on rationale quality by limiting information seeking during the task. This may occur if the participants already held a belief or attitude in relation to the topic. Although an attempt was made to minimise this effect by removing all participants with formal academic knowledge of Psychology, it would be difficult to fully control for all prior knowledge and beliefs which may influence the task. For the purposes of this research, being an exploration of the effects of self or other directed rationales based

arguments, the controls for prior knowledge and belief were considered sufficient to draw some interesting conclusions.

There could also arguably be an issue with the ecological validity of the rationale based argument elicitation, as the task was conducted in the absence of any feedback to the participant which may be common in face to face or synchronous collaboration with a co-learner. However, there are many instances in the real world context, such as individual e-learning environments and software design, for which people are required to construct standalone rationales without direct feedback. In this case the skill of argument and the need to argue well is of great importance for increasing understanding and the ability to competently construct effective arguments.

There is also the inherent problem of subjective analysis within the study, particularly for the quality ratings and manual structural analyses, which require subjective plausibility judgements on the part of the analyst. The Classical RST analysis may be somewhat weak in areas and caution needs to be taken as only one analyst undertook the full corpus analysis, with a small proportion of 10% of the rationales analysed by an independent rater for comparison. A way to overcome this in the future is to engage a number of consecutive analysts and ascertain the majority view or 'average' agreement.

Having said this, the findings from the automated parsers do help to shore up the findings of the Classical RST and other structural analyses. The Contrasts and Comparisons identified by the automated parsers (and the subsequent significant differences between the groups) do relate to the concepts of Counter Claims and Rebuttals (which provide the basis of the quality scale) as all three can be denoted by similar discourse markers. The marker 'but' is a common discourse marker and can signal any of these constructs so it is not surprising that they concur in terms of the differences between the groups. Previous research has indicated that these relations are the most difficult to differentiate between from both a human and computational approach. These possible parser limitations will be discussed in section 11.7.3.

Additionally, the analyses conducted involved a large number of comparisons, although an attempt was made to further refine the structural relations examined in the hypotheses. As the structural analysis approaches generated a large number of variables it is inevitable that a large number of statistical tests needed to be carried out. Particularly as these approaches are fairly novel in the application of argument analysis, it was deemed to reductive to only

examine one variable per hypothesis as the relations often encompass similar constructs, thus a grouping was considered more appropriate. The use of Bonferroni as a correction measure for studies with a large number of tests has been argued as too conservative (Bender & Lange, 1999). If this method were to be applied to the data in this investigation it may well be too conservative and thus result in the rejection of true hypotheses. For the purposes of this type of exploratory research that incorporates analysis methods with large number of variables, it was considered appropriate to restrict correction to the post-hoc analyses only, where applicable.

6.5 Conclusions and Further Work

6.5.1 Conclusions

Even with the limitations discussed above borne in mind, there are still some worthwhile conclusions that can be drawn from the evidence and logical next steps that can be taken in modelling rationale style arguments and producing adapted frameworks for rationale analysis.

The overall aim of this study was to elicit standalone rationales in a richer form to fully examine the possible structural differences between perceived self and other directed explanations. The work also intended to extend the previous findings that perceived other directed explanations may influence decision confidence and possess measurable differences in structural qualities. The application of several methodologies in the analysis revealed the relationships between these, as well as a relatively novel application of RST to rationale style arguments.

This study supported previous research regarding explanation in an educational context, by demonstrating that constructing a rationale can aid with the encoding of new information (with the use of a pre and post-test of new domain knowledge), and that the additional intervention of asking participants to explain for others may have the potential to enhance this effect. Although recall of new information is not a complete picture of learning, it is a part of the learning puzzle and important in facilitating understanding of procedural and conceptual information across all domains.

The findings also indicated that constructing rationales in the presence of a directional prompt intended to influence the author's perception, resulted in some intriguing

differences with regard to the argumentative structures produced and decision confidence. The OD group rated their decision confidence as higher on average than the SD group and the OD group rationales contained significantly more complex and higher quality argument relations. It is also clear from the findings that both self and other directed explanation occurred in both groups regardless of prompting. This resulted in the mean ratings for direction in the OD group being at the mid-point of the scale, therefore is more appropriate to label the perception of direction in this group as 'less self directed' as opposed to fully 'other directed.'

This finding, that a prompt for direction that suggests a future use for an argument to assist others, will be referred to as the 'perceived direction effect.' This control of the perception of direction is much easier to implement into a learning environment and more time and resource efficient than organising direct interaction, that is, if the goal of the activity is to enhance confidence and individual argument quality.

This is intriguing as the interaction and feedback level were kept consistent throughout both conditions suggesting that participants were naturally varied in their attitudes towards interactivity during the task. Kuhn & Reiser (2005) proposed that in order to become proficient in constructing defence within arguments, students need to have a visible tangible audience. However this research has shown that this is not necessarily the case and that a perception of interaction alone appears to have a comparable effect of triggering defensive arguments regardless of the physical context.

It is important to note that although all three RST analysis approaches identified varying types and numbers of relations due to the different methods of classification of rhetorical relations, they generally identified semantically and functionally similar constructs. The reasons behind this may lie in the fundamental rules used within the automated parsers to label relations. It is unsurprising that they differ to some extent and there remains a need for more consistent analysis methods. It is reassuring that there are differences between the automated parser results, as this indicates that differences between the Classical and automated parsers may not just be a case of a discrepancy or lack of consistency within the human and automated approaches. In addition, the identification of similar constructs within each tool and the Classical analysis (which largely correlate with one another) gives some confidence that the relations do hold and can be used as a basis for a future modelling of other and self-directed rationales. The correlations between the Classical RST findings

and the other frameworks are discussed in part four of the thesis and new adapted frameworks are proposed as a result.

6.5.2 Modelling the Findings

The findings thus far have given rise to an opportunity for modelling some of the influences and possible internal processes that may occur as a result of rationale style argument construction, as a function of the perceived direction effect. The model proposed here is primarily intended to summarise and illustrate the findings.

The model (outlined in Figure 20) is a high-level construct at this stage, mapping the overall argument structures present and the constraints on argument quality. The model incorporates the Toulmin elements of Backing and Rebuttal to broadly categorise one sided and two-sided arguments. Ideally, the findings will offer insight into modelling other types of arguments and provide suggestions for the constraints - such as the perception of direction - that may impact argument quality in terms of the use of 'Argue' type strategies.

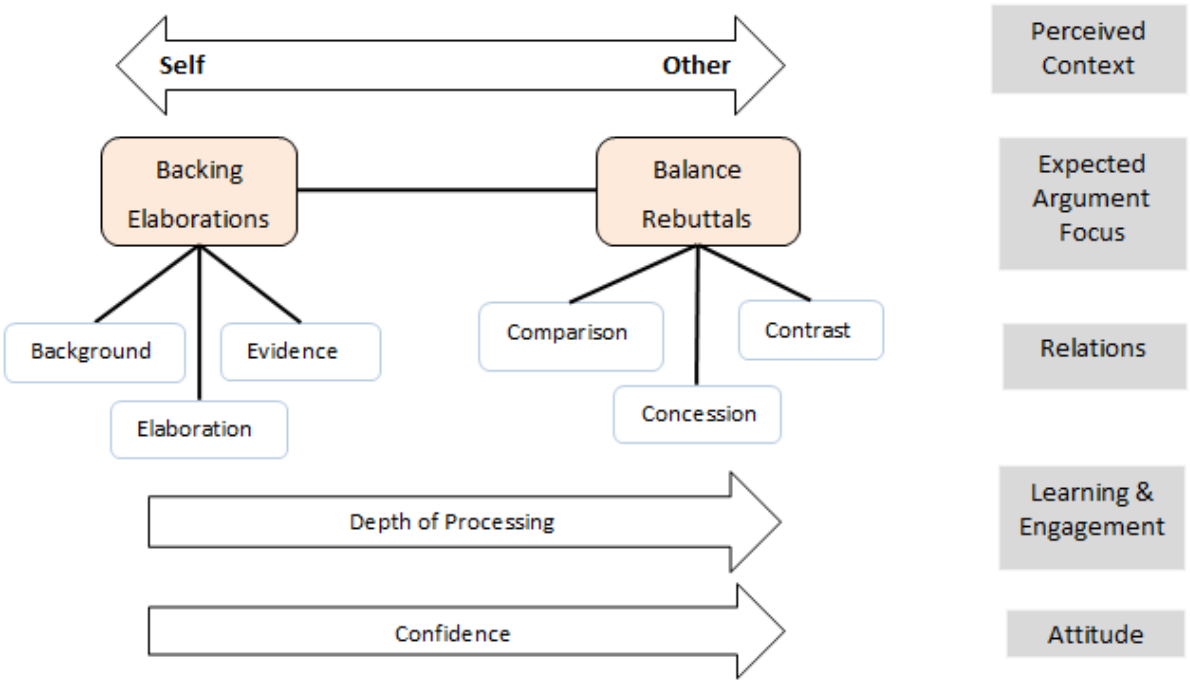

Figure 20 Model of the Perceived Direction Effect on Argument Structure, Quality and Attitude.

The previous model, developed on the basis of the exploratory study findings, suggested that the participants were possibly constructing arguments that were more balanced and

coherent as a result of being told using a directional prompt that their arguments may be used in the future to assist others. There also appeared to be a confidence effect between the groups. This basic representation of argument quality as a function of length and coherence can now be seen in this model and has been extended to incorporate the findings from this investigation. This model provides a richer view of argument in terms of more detailed structures and helps to visualise the effects. The model suggests that that explaining for another (or with the perception that another will use the rationale) prompts a different argumentative strategy. This has implications for findings in the Self explanation literature in that it suggests that the 'self explanation effect' may be confounded by a 'perceived others' effect.

In addition, it provides a framework for the scope of influence on rationale style argument structure. These structures pertain to either a one sided supportive approach in the sense of Backing for an argument or a two sided balanced approach, concerned with Rebuttals ('expected argument focus'). The one sided approach appears to consist of Background and Elaboration weighted arguments, whereas a two-sided approach can be seen to comprise of higher frequencies of Contrasts and Concessions. The one sided arguments appear to be more prevalent in self directed arguments and two-sided strategies are more prevalent as part of a less self directed approach.

The model suggests that the shift from a self to other directed approach is linked to an increase in the interaction with the task material (hence a learning effect) and an increase in positive regard for the decision made and the rationale produced. Additionally, this depth of processing may also be influenced by the perception of possible future interaction with an argument. This perception may trigger the use of more complex argumentative strategies which involve greater integration with the materials and hence greater chance of recall. Specific argument structures are indeed inherently linked to argument quality and the facilitation of a positive attitude towards the argument itself and the decision made on the basis of it. The model also extends these observations to suggest that argument quality may have an impact on the depth of processing of the materials and resources available at the time of constructing the argument and thus may facilitate retention and recall of information at a later date.

This effect is noteworthy at this stage; however, the recall performance test was relatively small and focussed on generic factual information from the task material. Additionally, it may be that some participants paid more attention to one side of the material over the

other, in essence ignoring information that did not concur with their position. This is a common human tendency and may be revealed if all participants were tested fully for recall of each item. Even so, in order to fully assess the impact on retention, a full test of all items may need to be conducted at a later date to assess long term retention. It must also be made clear that the 'depth of processing' effect outlined in the model refers to the superficial retention of factual knowledge and not an implication of the level of understanding of the material presented.

The theory of constructivist learning would suggest that learning occurs when interacting with another (Osborne, et al., 2004a). The findings from this study suggest that perhaps an internal dialogue is responsible for the increased learning and positive attitude towards the task, as demonstrated in the OD group. An inner dialogue in the absence of actual interaction may evaluate knowledge in a similar fashion to conversing with another person. It may be that an inner dialogue appears spontaneously for any learning context but that there needs to be an internalised belief that others may be or are present for the belief to take hold and thus improved strategies to emerge as a result. In other words, for this study the cue that the rationales would be used at a later date by another needed to be attended to and internalised. This did not always occur, as some participants stated the direction of their rationale was contrary to the group prompts. A full understanding of the impact of rationale construction on understanding versus the recall of declarative knowledge is currently outside the scope of this investigation but would prove interesting for future work.

The findings have implications for comparisons of self and other directed explanations, as self-explanations are generally considered to be a 'reflective' activity, which include information production, but few evaluative approaches. In contrast, other directed explanations may contain more elements to justify and evaluate elements within the arguments. The findings support observations by Xiao (2013b) that reflective reasoning styles tend to comprise of subject matter relations. As reasoning becomes more interactive, presentational relations are more prevalent.

It may be that the construction of other directed arguments fosters a more critical approach to the task and increases analytical behaviour. Critical thinking involves evaluation, developing questions, making links and developing arguments to support a position and a consideration of alternative views (Scoufis & Writing, 1999). If critical thinking is the task goal then the increase in Contrasts and evaluative approaches that the other directed

perception can trigger could have utility here. This is particularly interesting if this effect can still be measurable in the absence of any direct interaction or feedback and even when the audience is merely implied.

6.5.3 Next steps

The findings generated a wealth of data gathered using various frameworks. The revelation that complex argument structures were more prevalent in the other directed rationales prompted an interest in examining expert rationales in comparison to the novice rationales collected in this study. It may be that those who have a higher level of argumentative competency construct structurally comparable rationales to those who have constructed other directed rationales in this investigation. This may indicate that similar internal processes are responsible for these reasoning styles, even though the level of competency differs. It may suggest that those who are made aware that their arguments will be utilised in the future, may adopt 'expert' type strategies in their arguments.

7 Expert Versus Novice Rationales: A Comparison of Reasoning Styles

7.1 Introduction

The revelation that complex argument structures were more prevalent in the other directed rationales prompted an interest in examining rationales constructed by experts, to provide a comparison with rationales constructed by more novice arguers.

In order to fully examine the potential differences that may occur in the rationales, the literature comparing expert and novice composition may offer some additional considerations. In terms of argumentative writing, the literature tends to lean towards ascertaining the differences in the persuasive power of expert and novice writers. A persuasive discourse is a text produced to increase adherence to or positive regard of an audience to a concept, idea or object (Crammond, 1998). Persuasive writing, it could be argued, needs to be balanced, in other words, the main aspect of a persuasive piece is how successfully it counters possible arguments before they occur. This would certainly help to assist in avoiding a circular argument. This is particularly pertinent in this research as it appears that those cued to perceive future interaction with their arguments are writing more persuasively in this respect. However, the most valid persuasive measure of an argument lies in the impact on the receiver independent of author intention and it is clear that it may not always be a balanced argument that is more persuasive (Petty & Cacioppo, 1984).

Crammond (1998) examined essay style arguments constructed by student and expert writers. The work was intended to examine structural differences in arguments that were not a function of text length but argument density within the text. The study adopted a Toulmin model based analysis and confirmed that expert arguments appeared to contain more argumentative structures per text, including a higher frequency of rebuttals and more qualifying statements.

A further study of adolescent argument skills found that speaking aloud seems to promote better and potentially more persuasive arguments than writing alone (Felton & Herko, 2004). This effect may be due to the familiarity with spoken argument and the interaction that speaking aloud facilitates may provide more prompts to elicit complex argumentative structures. Bereiter and Scardamalia (1987) observed that novice arguers tend to adopt a

knowledge telling strategy that usually consists of a simple chain of claims. In contrast, more expert writers were thought to adopt knowledge transforming strategies.
In a rhetorical sense a transformation approach could be the use of a Concession, which involves the comparison and use of available information to highlight a flaw in an argument. Additionally, it may reflect a contrasting or comparison of concepts in order to demonstrate alternative viewpoints have been considered and assessed. Knowledge telling strategies could be those relations that pertain to statements of fact alone. In this investigation the Argue and State categories of relations will be considered as potentially knowledge transforming and telling strategies respectively.

In order to produce effective and (and possibly persuasive text), writers need to be aware of potential objections and defend possible counterarguments to their position. In essence, they need to be able to construct an inner dialogue that offers the opportunity to imagine the argument flow. This inner dialogue can inform the externalisation of the argument. In order for novice arguers to be effective they also need to have access to or be familiar with the opposing side of the argument.

To mitigate this, both the Expert and Novice groups in the current study were given access to the balanced information brief from which they could extract backing information and supporting statements for their claims. The presence of both sides of the argument in the brief may even act as a prompt to the participants inferring that they are expected to use information from both sides within their rationales. However, the findings from the previous work seem to suggest that many participants still constructed one sided arguments. Novice arguers may be overwhelmed by the cognitive demands of structuring writing and engaging in a balanced style of thinking, therefore the externalised argument is much weaker (Crowhurst, 1996).

It was considered worthwhile to examine what additional reasoning styles could be uncovered in expert arguments in comparison to novice arguments and whether this has congruence with previous work. The reasoning styles of experts may also reveal similarities to the trends detected in arguments as a result of a perceived direction effect.

7.1.1 Research Questions

This investigation aims to address the following research question:

1. Do expert and novice authors differ in their attitude and approaches to rationale construction in terms of the use of knowledge telling and knowledge transforming strategies and are these reflected in the quality and use of measurable argumentative strategies?

7.1.2 Hypotheses

In order to address the research question based on the research discussed and the previous work, the following hypotheses are proposed:

H1: Expert authors will report higher levels of confidence in their decision in comparison to novice authors.

H2: Expert authors will report higher ratings of perceived persuasiveness for their rationales than Novice authors.

H3: Expert authors will construct arguments with more instances of Rebuttals and Counter Claims than Novice authors, as measured by a Toulmin analysis and will be confirmed by a significant difference in assigned quality scores between the groups.

H4: Expert authors will have more instances of Argue type relations, (indicative of knowledge transforming strategies) specifically Contrasts, Comparisons and Concessions (as shown by the Classical RST, automated PDTB and HILDA parser analyses) compared to Novice authors.

H5: Novice authors will have more instances of State type relations (indicative of more knowledge telling strategies) compared to Expert authors. This will be assessed by Classical RST and automated PDTB and HILDA parser analyses.

7.2 Method

7.2.1 Participants

A total of 18 participants comprised the Novice argument group. The selection process for this group is described in the procedure section. A new sample of 18 Expert rationales was then gathered for comparison. The Novice group comprised of 10 males and 8 females (Median age = 21 (M = 22.7). The Expert group comprised of 7 males and 11 females (Median age = 22.5 (M = 24).

7.2.2 Design

7.2.2.1 Independent Variables

A between subjects design was used with the level of subject and academic argument expertise as the independent variable. The two levels are described below:

- **Novice** – This group comprises undergraduate students with no experience with or formal academic study of a social science subject.

- **Expert -** This group comprises professionals in the field of the social sciences.

7.2.2.2 Dependent Variables:

The full list of dependent variables can be seen in *Table 34*. A full explanation of the analysis methods for these variables can be found in the previous investigation; section 6.2.3.5.

	Dependent Variable	Specific Measurement
1	Decision Confidence	Likert rating scale of 1-7
2	Perceived Persuasiveness of Rationale	Likert rating scale of 1-7
4	Quality of Rationale	Five level Quality Framework
6	Toulmin Elements	Number of Rebuttals and Counter Claims
7	Argue Type Relations	Contrast (Classical RST and PDTB, HILDA) Concession (HILDA & RST) Comparison (HILDA) relations.
8	State Type Relations	Conjunction (Classical RST and PDTB) Restatement (Classical RST and PDTB), Background (Classical RST and HILDA), Elaboration (Classical RST and HILDA) Justify (Classical RST) Attribution (HILDA)

Table 34 Summary of dependent variables: attitude, structure and quality.

7.2.3 Procedure

The participants in the previous study were undergraduate students with no experience with or formal academic study of a social science subject. These participants were considered 'novice' both in their knowledge of Psychology and their experience with professional debate and argument, such as Journal article composition. The Expert rationales appeared lengthier than the previous study, with a mean of 159 words compared to 67 in the SD group and 87 in the OD group.

7.2.3.1 Novice Group Selection

To ensure a fair comparison was conducted the novice sample for the current study was created by selecting the nine lengthiest rationales from both the previous OD and SD groups. This was to ensure that the rationale lengths for the Novice group were comparable and not significantly different to the Expert group (this will be confirmed in the findings). This will assist in drawing reliable comparisons between the frequencies and types of relations used

in the Expert and Novice groups, which will be more likely to be a function of argumentative competency and not simply a result of lengthier rationales.

7.2.3.2 Expert Group Selection

The expert comparison sample was gathered on the basis that they were professionals in their field who regularly compose argumentative texts (such as critical reviews, journal papers). The 'Experts' are not primarily experts in argument per se, but have more extensive expertise and knowledge in the field of psychology, having studied at a minimum of undergraduate level. Therefore these 'Experts' would be expected to be competent at arguing in a social science context, in comparison to those participants in the second study who had not formally studied psychology and were still studying at undergraduate level.

The Experts were given the same information brief as the Novice decision makers from the previous study. The Expert group were asked to decide if they thought that humans were innately violent and to construct a rationale for their position using the information provided. The task was administered on an individual basis via email. They were instructed to compose their rationales in natural free text and were not instructed to write for any audience in particular. The same decision evaluation measure questions utilised in the previous study were administered to the Expert sample. The questions ascertained the level of decision confidence, whether or not the rationale was considered persuasive by the author and whether the rationale was self or other directed.

7.2.3.3 Structural Analyses

All Expert rationales were subjected to the Toulmin elements analysis, the Classical RST analysis and the Toulmin based quality scheme analysis. In addition, the automated parsers HILDA and the PDTB were used to analyse the text.

As the Alternative and Instantiation relations were very infrequent in the previous work these relations were excluded from the Argue and State analyses for the PDTB parser (see Appendix, Table 64). Additionally, the Summary and Restatement relations were very infrequent in the previous Classical RST analysis and were also excluded from the analysis. The Argue category of relations for the Classical RST analysis was further refined prior to analysis to focus on the Contrast and Concession relations as these are present in the automated approaches and appear to be a key variable according to the previous

investigations (see section 5.3.2.4 and 6.3.8).Thus the hypotheses for this investigation focus on a slightly more refined and specific set of relations.

7.3 Findings

Any relation or argument element that numbered a total of zero in the entire sample for either group was removed from any further statistical investigation. The data was treated as non-parametric and thus a series of Mann Whitney U and non–parametric correlational analyses were carried out where appropriate to determine statistical significance. All variables were also tested using an independent t-test analysis, and the findings concurred with the non-parametric approach. However, the low sample size and frequency of zero scores for many relations resulted in unequal variances for the many of the relation categories across the two groups.

7.3.1 Rationale Length

Firstly, to confirm that there was not a significant difference in average rationale length between the Expert and Novice groups a statistical analysis was conducted. As the Novice group was approximately matched to the Expert group there was no significant difference found between the word counts (U = 160, p = .963, Z = -.063). The mean word count for the rationales within each group are summarised in Table 35.

	Rationale Length (words)	
	Mean	***SD***
Expert	159.00	65.29
Novice	146.61	33.90

Table 35 Rationale length: Expert and Novice groups

7.3.2 Hypothesis 1: Confidence Comparison – Expert and Novice

The average ratings for the each of the decision measures in both the Expert and Novice groups is summarised in Table 36.

	Expert			Novice		
Evaluation Question	**Mean**	***SD***	**Mode (Median)**	**Mean**	***SD***	**Mode (Median)**
Confidence	5.89	1.08	6 (6)	5.83	1.20	6 (6)
Persuasive	5.39	1.20	6 (6)	4.72	1.41	5 (5)
Self/Other	5.06	1.86	7 (5)	3.33	1.64	4 (3.5)

Table 36 Average decision evaluation ratings for Expert and Novice groups.

The first hypothesis states that the Expert group will report higher confidence ratings than that Novice group. The mean ratings for each of the decision evaluation measures are summarised visually in Figure 21. The average confidence ratings for both of the Expert and Novice groups are almost impossible to differentiate between and no significant difference was found (U = 161.5, p = .988, Z = -.017). In contrast, the Expert groups appear to consistently rate the other decision evaluation measures as noticeably higher than the Novice groups.

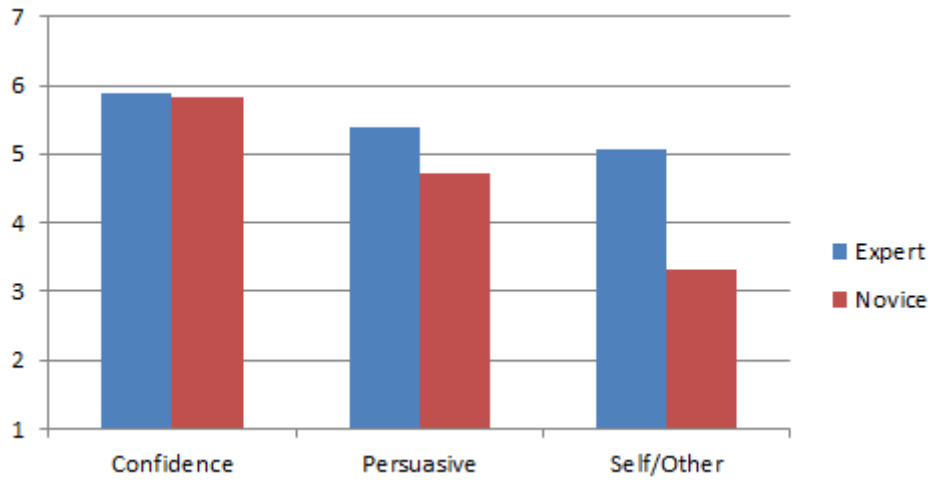

Figure 21 Means (y-axis) for decision evaluation measures (x-axis) in both groups.

7.3.3 Hypothesis 2: Persuasiveness Comparison – Expert and Novice

The second hypothesis states that the Expert authors will rate their rationales as higher in perceived persuasiveness than the Novice group. Figure 21 suggests that the Experts appeared to view their rationales as generally more persuasive. This difference, although marked, was not statistically significant (U = 116.5, p = .152, Z = -1.499).

7.3.3.1 Directional Ratings Between Groups

Experts also seemed to rate their rationales as significantly more other directed in comparison to the Novice group (U=83.5, p = .011, Z= -2.528, r = .44). However no prompts were given to suggest future use of the rationales. The Expert ratings were compared to the Novice sample which was constructed using nine rationales from both the self-directed and other directed groups in the previous study. This was intended to attenuate the directional ratings to be more neutral by using rationales from both conditions. However, as the Expert sample received no such manipulation of audience perception it may not be appropriate carry out a direct comparison for this particular variable.

7.3.4 Hypothesis 3: Toulmin Element and Quality Comparison – Expert and Novice

The Toulmin argument structures were identified in both the Expert and Novice samples. The average number of each of the elements within each group is summarised in Table 37. Figure 22 clearly shows that the Expert group appear to have a preference for Rebuttals and Counter Claims compared to the Novice group. In contrast, the Novice group appear to reply more heavily on Backing elements within their arguments. There does not appear to be a discernible difference in the number of claims made in the arguments between the groups. This is most likely due to the arguments being generated in response to a question that required a positon to be taken. Thus the claims are more likely to be just one statement of position for either side, followed by the rationale and supporting statements.

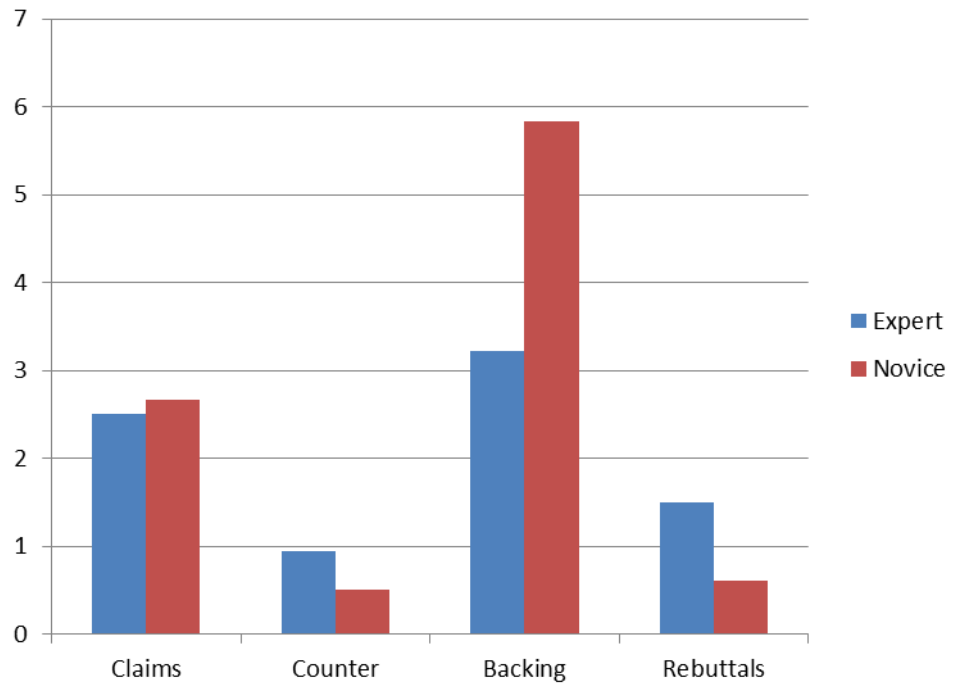

Figure 22 Means (y-axis) for Toulmin model elements (x-axis) within Expert and Novice groups.

Toulmin Element	**Group**	**Mean**	***SD***	**Median**
Claims	Expert	2.50	0.92	2
	Novice	2.67	1.19	3
Counter Claims	Expert	0.94	0.64	1
	Novice	0.50	0.51	0.50
Backing	Expert	3.22	1.99	3
	Novice	5.83	3.67	5
Rebuttal	Expert	1.50	0.62	1
	Novice	0.61	0.61	1

Table 37 Descriptive statistics for Toulmin element analysis for Expert and Novice groups.

The Expert group had significantly more Counter Claims (U=103.5, p =.036, Z = -2.094) and Rebuttals (U=58.5, p = .000, Z = -3.591, r = .59) than the Novice group.

The Expert and Novice groups also differed significantly in the quality scores assigned for the rationales (U = 99, p = .047, Z = -2.145, r = .37). The statistics for the quality scoring in

both groups can be seen in Table 38. The Expert group rationales were rated as significantly higher quality overall.

	Quality Score		
	Mean	***SD***	**Median**
Expert	4.33	0.69	4
Novice	3.67	0.97	4

Table 38 Quality score descriptive statistics: Expert and Novice groups

7.3.4.1 Post Hoc Analysis: Backing Elements

As the rationales were similar in length it was considered worthwhile to examine whether Backing comprised significantly more of the argumentative structures in the Novice group compared to Expert group. The Novice group did in fact appear to have a significantly higher mean number of Backing elements (U=71, p = .003, Z = -2.948, r = .40) within their arguments. This finding concurs with the literature expectations that novice arguers will tend to adopt a knowledge telling strategy and experts adopt a knowledge transforming strategy.

7.3.5 Hypothesis 4: Argue Relation Category Comparison – Expert and Novice

The fourth hypothesis states that expert authors will have more instances of Argue type relations than the novice authors.

7.3.5.1 Classical RST Argue Relation Analysis

The Argue category of relations for the Classical RST was refined to focus on the Contrast and Concession relations as these are present in the automated approaches. The mean number of each relation within the rationales are shown in Table 39 and represented visually in Figure 23. The full summary table can be found in Appendix 13, Table 30. Primarily, there appear to be significantly more instances of Classical RST Contrast relations (U=51.5, p = .000, Z = -4.021, r = .60), in the Expert group. There was no significant difference in the number of Concession relations in the Classical RST analysis (U = 159.5, p = .938, Z = -.085).

Argue Relation	Group	Mean	SD	Median
Concession	Expert	1.39	0.70	1.5
	Novice	1.56	1.20	1
Contrast	Expert	1.11	0.96	1
	Novice	0.06	0.24	0

Table 39 Means for Classical RST Argue relations in the Expert and Novice groups.

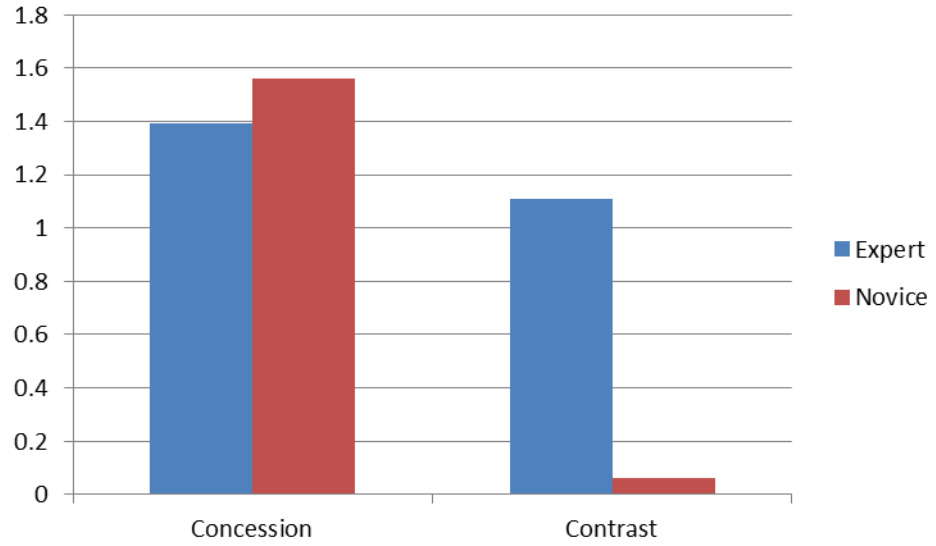

Figure 23 Means (y-axis) for Classical Argue RST Argue relations (x-axis) within Expert and Novice groups.

7.3.5.2 HILDA Parser Argue Relation Analysis

The types and means for the Argue category relations identified by the HILDA parser within each group are summarised in Table 40 and visually in Figure 24. The full summary table is available in Appendix 13, Table 31. Again, any relations which did not feature in one or more of the groups were removed from the statistical analysis.

The Expert rationales appeared to contain significantly more Contrast relations (U=98.5, p =.031, Z = -2.158, r =.38) than the Novice rationales. No significant difference was found between the groups for the Comparison relations (U = 162, p = 1.000, Z = .000).

Relation	Group	Mean	*SD*	Median
Comparison	Expert	0.06	0.24	0
	Novice	0.06	0.24	0
Contrast	Expert	1.17	0.99	1
	Novice	0.50	0.62	0

Table 40 Means for HILDA Parser Argue relations in the Expert and Novice groups.

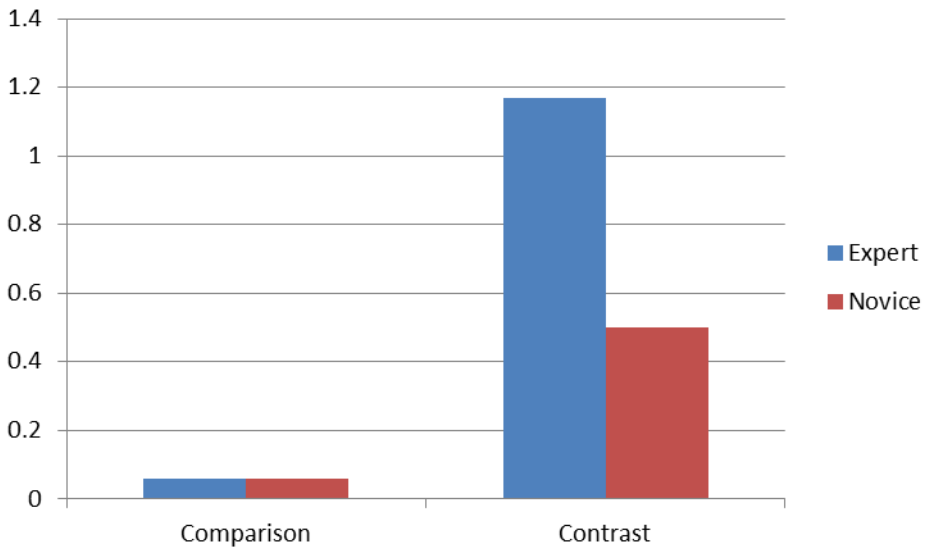

Figure 24 Means (y-axis) for HILDA Parser Argue relations (x-axis) within Expert and Novice groups.

7.3.5.3 PDTB Parser Argue Relation Analysis

The average number of each Argue type relation detected by the PDTB parser, for each group, is summarised in Table 41 and visually in Figure 25 (the full summary table is available in Appendix 13, Table 32). Again, any relations which did not feature in one or more of the groups were removed from the statistical analysis with the Alternative relation removed due to its infrequency in the previous findings. No Concession relations were found in the Expert group.

Relation	Group	Mean	*SD*	Median
Concession	Expert	0	0	0
	Novice	0.22	0.15	0
Contrast	Expert	2.06	1.11	2
	Novice	1.33	1.28	1

Table 41 Means for PDTB Parser Argue relations in the Expert and Novice groups.

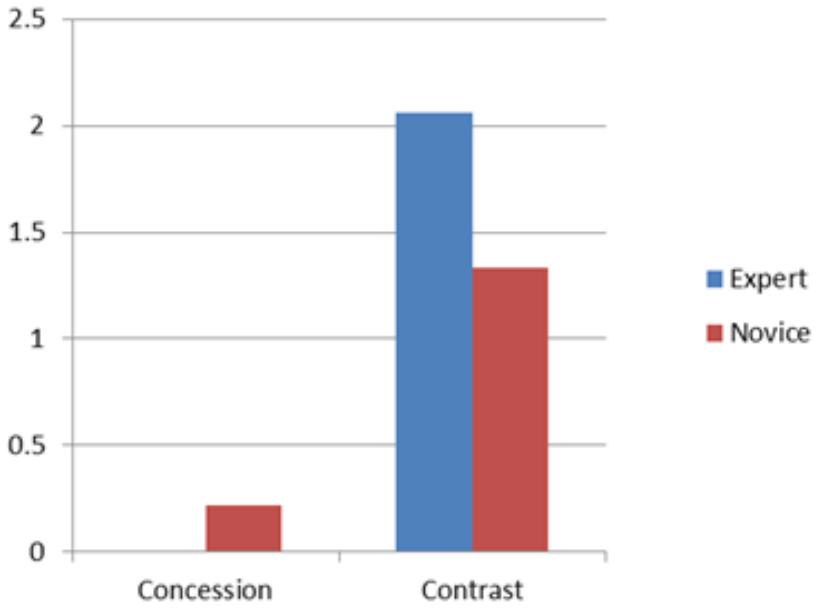

Figure 25 Means (y-axis) for PDTB parser Argue relations (x-axis) within Expert and Novice groups.

Figure 25 clearly shows a marked difference in the mean number of Contrast relations between the groups. A Mann Whitney U analysis indicated that a significant difference exists between the use of Contrast relations (U= 97, p = .030, Z = -2.172, r = .61), with the Expert group once again, containing more of these constructs.

7.3.6 Hypothesis 5: State Relation Category Comparison – Expert and Novice

To examine whether the Novice groups constructed rationales with higher number of State category relations the groups were compared using the three approaches.

7.3.6.1 Classical RST State Relation Analysis

The average number of each State type relations found using the Classical RST analysis for each group is summarised in Table 42. The means for the relations are represented visually in Figure 26, with the full summary table of descriptive data available in Appendix 13, Table 32. The Summary and Restatement relations were excluded from consideration due to the infrequency of these relations in the previous research.

Relation	Group	Mean	*SD*	Median
Elaboration	Expert	1.50	1.20	1.5
	Novice	3.67	2.20	3.5
Conjunction	Expert	0.72	0.75	1
	Novice	0.22	0.65	0
Justify	Expert	0.78	0.73	1
	Novice	0.28	0.46	0

Table 42 Descriptive statistics for Classical RST State relations in the Expert and Novice groups.

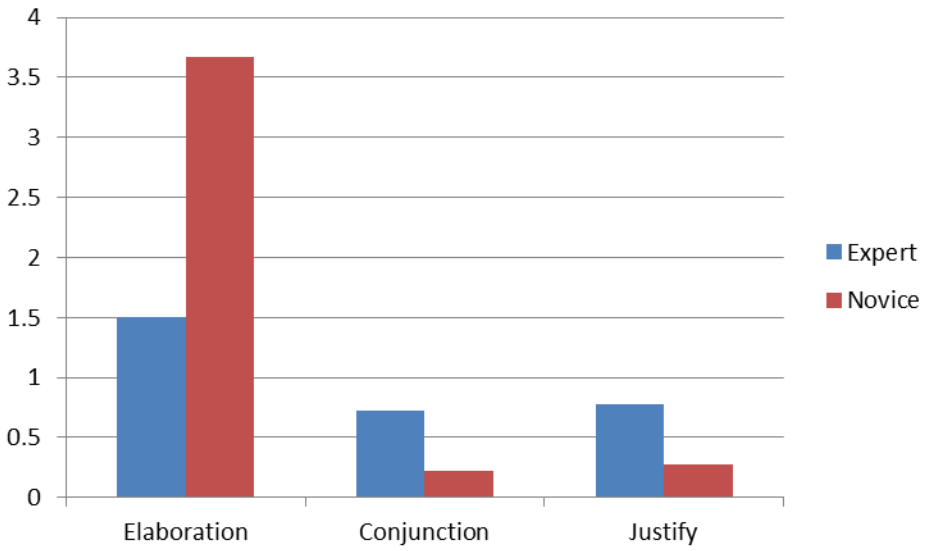

Figure 26 Means (y-axis) for Classical RST State relations (x-axis) within Expert and Novice groups.

No significant differences were found between the groups in the use of Summary (U= 157, p = .888, Z = -.230) and Restatement relations (U = 126, p = .265, Z = -1.764). A significant difference was found between the groups for the Elaboration relation. The Novice group contained significantly more of these relations than the Expert group (U = 69.5, p = .003, Z = -2.978, r = .52).

Contrary to the hypothesis, the two variables of Conjunction (U = 97, p = .040, Z = -2.468) and Justify (U = 96.5, p = .037, Z = -2,368) relations were actually more frequent in the Expert group.

7.3.6.2 HILDA Parser State Relation Analysis

The descriptive statistics for the State category relations detected by the HILDA parser can be seen in Table 43 and the means represented visually in Figure 27. The Summary relation was not detected in either of the group's rationales, and was not detected in the previous experimental work using the parser (Table 65). As the PDTB parser detects the Restatement relation (see Table 44), and Summary was recognised in the manual Classical RST analysis in the previous study (Table 62), the HILDA parser's lack of detection for this relation is possibly a functional error.

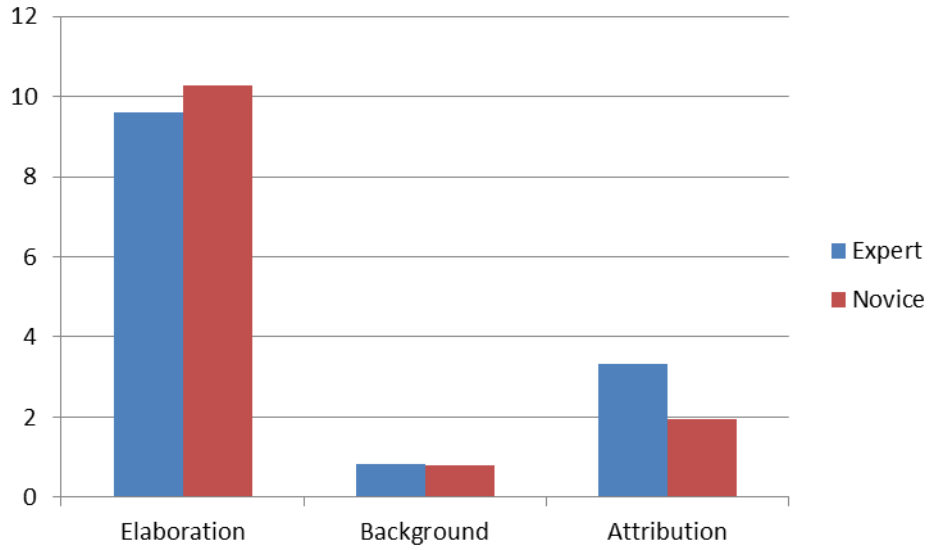

Figure 27 Means (y-axis) for HILDA Parser State relations (x-axis) within Expert and Novice groups.

As can be seen in Figure 27 there is a slightly more extensive use of Elaboration relations in the Novice group. However, no significant differences were found in the use of the Elaboration (U=139.5, p = .481, Z = -.715) or Background (U = 159.9, p = .932, Z = -.085) relations between the groups.

Relation	**Group**	**Mean**	***SD***	**Median**
Elaboration	Expert	9.61	4.95	9
	Novice	10.28	3.68	9.5
Background	Expert	0.83	0.99	0.5
	Novice	0.78	0.73	1
Attribution	Expert	3.33	1.60	3.5
	Novice	1.94	1.26	2

Table 43 Means for HILDA Parser State relations in the Expert and Novice groups.

Again, somewhat contrary to expectations raised by the hypothesis, the Expert and Novice group also differed in the number of Attribution relations, with the Expert group containing significantly more of these (U = 76, p = .006, Z = -2.773, r = .43) than the Novice group.

7.3.6.3 PDTB Parser State Relation Analysis

The descriptive statistics for the State category relations detected by the PDTB parser can be seen in Table 44 and visually in Figure 28.

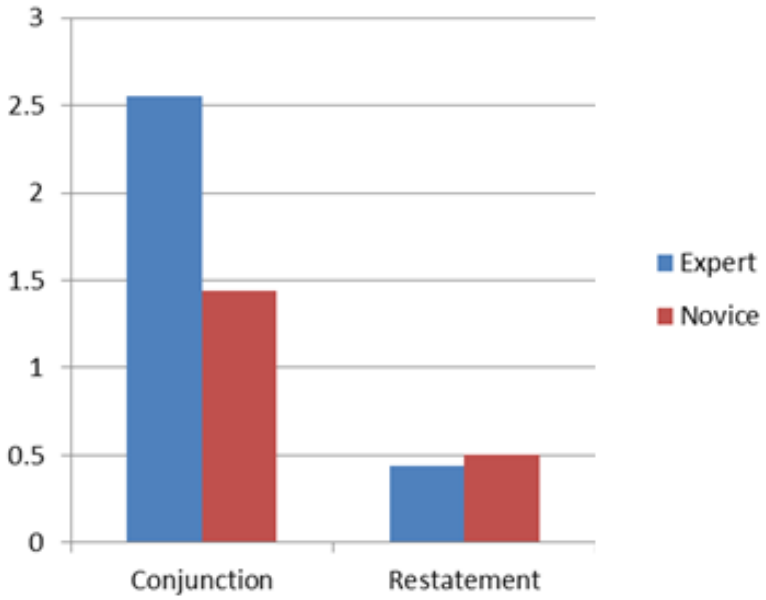

Figure 28 Means (y-axis) for PDTB parser State relations (x-axis) within Expert and Novice groups.

Relation	**Group**	**Mean**	***SD***	**Median**
Conjunction	Expert	2.55	3.15	2
	Novice	1.44	1.15	1
Restatement	Expert	0.44	0.71	0
	Novice	0.50	0.79	0

Table 44 Means for PDTB Parser State relations in the Expert and Novice groups.

Although there appears, in Table 44, to be a differences in the use of Conjunction relations, no statistically significant differences were found between the groups in the use of either Conjunction (U = 114.5, p = .123, Z = -1.543) or Restatement (U = 161.5, p = .985, Z = -.018) PDTB relations.

7.4 Discussion

The discussion will address the findings in terms of support for the hypotheses proposed and the implications for the previous work.

7.4.1 Hypothesis 1: Comparison of Confidence Between Groups

The findings did not indicate a significant difference in reported confidence between the groups. Thus the null hypothesis is accepted in this instance. This may be explained by the similar rationale lengths between the groups. As the Novice group was selected on the basis that the average word length would match the Expert group, previous research would suggest that if confidence is linked to effort (Sieck & Yates, 1997), it could be assumed that the levels of confidence would be similar in this case.

7.4.2 Hypothesis 2: Comparison of Persuasiveness Between Groups

The second hypothesis stated that Expert authors will report higher ratings of perceived persuasiveness for their rationales than Novice authors. The null hypothesis is supported in this case as no significance difference was found between the groups for the perception held by the authors of how persuasive the arguments were. A possible explanation for this finding would again be related to the effort-performance belief effect. The similar word lengths suggest comparable effort was exerted in both groups to construct the arguments. Therefore, if the perception of how persuasive your arguments are is a measure of how good you think they are, it follows that this may also be influenced by the length of the argument. It may also be that Novice arguers are not competent in recognising which aspects of an argument (such as rebuttals) may be more convincing to another, therefore their assumption of the impact of their argument is not based on specific structure, whereas an Expert arguer (whose argument may contain more rebuttals) may recognise this. This explanation requires investigation to ascertain if the author attitudes are based on specific structural features, such as the presence of rebuttals, and whether the level of expertise determines your ability to recognise the importance of these within an argument.

7.4.3 Hypothesis 3: Between Groups Toulmin Elements and Quality Comparison

The third hypothesis stated that Expert authors will construct arguments with more instances of Rebuttals and Counter Claims and will therefore be of higher quality than Novice authors. This hypothesis is supported by the findings that the Expert group rationales contained significantly more Counter Claims and Rebuttals than the Novice group. This difference in strategy is also mirrored in the Toulmin based quality ratings for which the Expert group outperformed the Novice group.

The use of rebuttals is considered a skilled strategy that is desirable in human argument to decrease the occurrence of circular arguments. It is evident that this is a skill held by the Expert group, whereas the Novice arguers appear to use these components of argument far less frequently.

In Novice arguments, an increased amount of text within the arguments was dedicated to backing (or evidence relations) elements, compared to Expert arguments. Although in previous research Warrants were considered as important aspects of an expert argument, this entirely depends on the context and the perception held by the author of the intended audience. Warrants are not always necessary; if the data is given as part of a task brief, there is not an explicit need to state warrants as both parties mutually agree (or it is assumed) that the information is accurate or accepted as true. For this reason, the examination of warrants was excluded from this study.

A post hoc analysis of the use of Toulmin Backing elements also confirmed that the Novice group displayed a significantly greater preference for these elements compared to the Expert group. These findings concur to some extent with the literature expectations that novice arguers will tend to adopt a knowledge telling strategy (shown in the use of Backing) and experts adopt a knowledge transforming strategy (shown in the use of Rebuttals).

7.4.4 Hypothesis 4: Argue Relation Category Comparison

The fourth hypothesis was concerned with the use of Argue relations and how this varied between groups. Expert authors were predicted to have more instances of the Argue type relations of Contrasts, Comparisons and Concessions. It was proposed that the Argue relations would be indicative of knowledge transforming strategies. The presence of Contrast relations was tested by the Classical RST, HILDA and PDTB parsers. Overall, the findings support the hypothesis by confirming that the Expert rationales contained more

Contrast (identified by the Classical RST, HILDA and PDTB parser) relations than the Novice arguments.

The Expert group appeared to contain no Concession relations detected by the PDTB parser. This is surprising, however, the findings from the second study (see section 6.3.8) and the parser definitions (Appendix 4) appear to suggest that the Concession relation in the PDTB parser may be labelled instead as a Contrast relation, as both form the Comparison class in the definitions. This assumption would be supported by the significant differences between the groups in the use of Contrast relations, which are detected in all three approaches, suggesting they are comparable. In addition, the distinction between the Contrast and Concession relations can be problematic as they rely on similar discourse markers. This issue is discussed alongside other parser limitations in section 11.7.3.

7.4.5 Hypothesis 5: State Category Relation Comparison

The final hypothesis, in contrast, was concerned with the use of State type relations and predicted that Novice authors would utilise more of these strategies than the Expert authors. The State category of relations was considered to be representative of knowledge telling strategies and therefore perhaps less complex than the Argue category in terms of effort and skill required to construct. This was assessed by a Classical RST analysis and both the automated PDTB and the HILDA parser.

Only one of the analyses appears to support the hypothesis but this is not consistent across the three analysis approaches and in some cases the findings are contrary to the expectations. The Novice group rationales were shown to contain significantly more Elaboration relations than the Expert group. The Novice group were not found to contain significantly higher numbers of any other State relations in any of the analysis methods.

The finding of increased Elaborations in the Novice group, although perhaps indicative of knowledge telling strategies, is contradictory to previous work that suggested that Elaborations were a cue to importance and persuasiveness (Rottman & keil, 2011) and thus indicated better arguments. It seems that the Expert writers preferred to adopt a balanced strategy when externalising their argument as opposed to opting for including a greater volume of elaborative type elements. It may be the case that the expert arguers are utilising an inner dialogue that mimics the effect of another person offering opposition, hence more counter argument strategies are employed.

The findings that proved to be contrary to the expectations raised by the fifth hypothesis include the discovery that the Expert groups contained more Classical RST State relations of Conjunction and Justify, as well as a higher prevalence of Attribution relations detected by the HILDA parser. Attribution relations were found to pertain to statements that contain 'suggest' and 'shows' as well as 'I think' which could suggest that the Expert rationales contained more backing and data which was attributable to a reputable source compared to the Novice rationales. This tendency may reflect the skill of academic writing in this context, as those who are familiar with the process will be proficient in citing reputable sources to support any claims. A habit that is possibly not as well ingrained in the novice arguers. Additionally, the Expert arguers may be more confident than the Novice arguers in citing their own opinion within their arguments due to their increased knowledge and skill and thus will use more instances of 'I' which will be detected as Attribution by the parser.

Both knowledge telling and knowledge transforming strategies could be considered useful depending on the context. If the objective of the activity is to learn conceptual knowledge, arguments with more Backing elements may be helpful. Although research has suggested that structures that require more critical ability and evaluation actually increase engagement with text rather than just simply reciting information to support a single claim. Expert rationales arguably contained more complex structures in terms of the competency required to form them; these processes have been linked with increasing and supporting critical ability. In comparison the Novice arguers appeared to prefer Backing elements in their arguments overall. However, the differences between the use of State relations in the groups is far less clear. If State relations are to be considered indicative of knowledge telling strategies is it then surprising that Expert arguers exceeded Novice arguers in the use of these in a number of cases. Overall, the consistent finding of Contrasts as a key Expert strategy, along with the related increase in quality and use of rebuttals, echoes some of the findings for the other directed group in the previous study (see section 6.3.7 and 6.3.8.3).

7.4.6 Implications for the Perceived Direction Effect

In terms of relating the findings from the Expert versus Novice examination to the effects of perceived self or other directed arguments it may be that the results have some bearing on drawing inferences about the cognitive processes at play. Those who construct other directed rationales may be drawing on cognitive processes similar to those who are competent at structuring written arguments. It may be that prompting to write in an other

directed manner, results in the triggering of an inner dialogue which guides the structure of the argument, and highlights the place for rebuttals and possible counter arguments. This may produce arguments and cognitive processes that would be more likely found in a direct interaction context.

The OD group investigated in chapter 6 and the Expert arguers in this study both produced superior quality arguments in terms of the significantly higher number of Contrast relations, and the use of Rebuttals (measured by the Toulmin and quality analysis). On the surface it appears that asking a person to justify a response prompts an automatic consideration of the possible direction or future use for the argument. This in turn prompts a strategy that influences argument structure which may be the use of an inner dialogue to help construct an argument. This is evidently an expert strategy regardless of the perceived direction of the argument.

Contrary to suggestions that in order to argue effectively you need an active tangible audience, it appears that increasing the perception that an argument will be viewed by others, regardless of actual audience, will elicit good 'expert' style arguments. It may be that that the Experts are engaging in an inner dialogue that prompts a knowledge transforming strategy, which in turn increases argument complexity. This is in contrast to the Novice arguers that appear to adopt a knowledge telling strategy that appears more concerned with paraphrasing or reciting relevant conceptual information.

7.4.7 Limitations

This sample size in this investigation was considerably smaller which may have influenced the findings. In addition only one analyst carried out the Classical RST and Toulmin analyses. The automated parsers also have limitations associated with them and the specifics of these will be discussed in the next chapters. Some of the constructs that may have been expected to be more common in Expert arguments such as Concessions were not apparent. This may be a result of sample size restrictions and errors in the Classical, or indeed, the automated parser analyses.

In terms of the methodology it needs to be acknowledged that the Expert authors carried out their task without a Skype connection to the investigator. As the Novice group sample consisted of rationales that were elicited via a black Skype video feed connection it would be a valid criticism to suggest that the task environments were not identical for both groups.

The Expert group were given the same instructional brief prior to completing the task via email that insists that the task be carried out alone in an environment free of distraction. However, the rationale elicitation task and information brief were identical for both groups and as such a comparison was still considered appropriate.

Additionally, due to the number of variables and comparisons conducted it is important to acknowledge that the use of Bonferroni corrections is recommended. However, this measure is extremely conservative and its application to exploratory investigations using large numbers of variables is often not advised due to the risk of rejecting a hypothesis when it may in fact be true. However, the consistent findings of Contrast relations systematically differing between groups through the first to the current investigation, does support the conclusions that the presence of these relations varies depending on directional prompting and expertise.

The next investigation extends the scope to examine how the externalised rationales can convey the author attitudes and a sense of competency through the arguments presented.

8 Rationale Evaluation: Exploratory Investigation

Rationale sharing in a computer supported environment is an emerging area of research. Sharing rationales appears to help collaborating individuals maintain control and monitor the quality of group work and resolve conflicts. Recent research conducted by (Xiao, 2013b) has demonstrated that sharing rationales within a group activity can affect processes such as grounding and activity awareness. The rationales may assist in making cognition of group members more transparent. Rationale sharing also appears to help promote awareness of the knowledge, contributions and expertise of other group members.

To demonstrate a small step in examining how rationales that are varied in terms of rhetorical structure impact upon those who receive them, a small study using a sample of the rationales was conducted. The investigation will help to give a perspective on how the findings thus far may relate to behaviour change, persuasion and perception of an author's attitude based on the arguments presented. The relationships between the attitudes towards the rationales and their structural properties are exploratory at this stage due to the small sample and limited range of rationale types used.

It is difficult to assess the persuasive power of rationales per se as this type of influence is largely reactive and many will have a pre-defined view on the particular controversial topic of human aggression. However, it was considered to be useful to examine how people perceived the quality, purpose and author intention within a sample of rationales, and to see whether the average assessments made regarding the rationales concurs with the original author's views and attitudes.

8.1 Design

8.1.1 Aim

The broad aim of this study is to examine whether the attitudes of the authors such as confidence held, persuasiveness and intended direction translate to the reader and whether the perceptions of the quality of the arguments received may be a function of the rhetorical structures present within.

8.1.2 Procedure

8.1.2.1 Participants

A total of 24 participants were asked to evaluate four rationales selected from the second study. The participants were selected on the basis that they had not formally studied a social science discipline. This was to ensure that judgements of argument quality were more reliably attributable to the structuring of the arguments, and less reliant on judgements about the accuracy of the data included.

8.1.2.2 Rationale Selection

The four rationale samples were chosen at random - from those constructed in the previous investigation in chapter 6 - that were between 95 and 110 words in length; half were selected from the OD directed group and half selected from the SD group. This was to ensure that the rationales contained varying structures but had comparable lengths. The rationales contained similar arguments; that human aggression is an interaction of both innate and environmental factors.

The rationales used can be seen in Appendix 15. The rationales differed in terms of their structure and content, as well as the perceptions held by the participants who constructed them. Rationale One takes a one sided position on the argument, whereas Rationale Two initially presents agreement with both nature and nurture positions and eventually concedes to an agreement with the nurture perspective. Rationale Three does not choose a polarised position and in a similar approach to Rationale Two, it contains arguments that indicate that either decision is plausible. The fourth rationale leans more to the nurture side of the debate but remains open minded. As such it would be expected that Rationale One and Four may be the most likely to be disagreed with as they confirm a choice, whereas Rationale Two and Three are possibly more difficult to disagree with as they essentially sit in the centre.

The specific structural features and reported author attitudes towards the individual rationales are summarised in Table 45.

		Rationale 1	Rationale 2	Rationale 3	Rationale 4
	Length (words)	106	104	110	97
	Persuasiveness	3	7	5	6
	Self Other	2	6	4	4
	Confidence	4	6	7	5
	Quality	3	4	3	3
Toulmin	Claims	2	3	2	3
	Counter	0	1	0	1
	Backing	3	4	7	2
Classical RST	Concession	0	2	0	1
	Evidence	4	3	3	3
	Justify	1	0	1	0
	Antithesis	3	0	0	0
	Interpretation	0	1	2	2
	Evaluation	1	0	1	0
PDTB	Contrast	0	2	0	1
	Cause	5	2	0	1
	Instantiation	0	0	1	0
	Asynchronous	0	0	2	0
	Synchronous	0	1	1	1
HILDA	Elaboration	9	7	11	8
	Attribution	0	6	1	3
	Contrast	0	1	0	1
	Explanation	1	0	0	0
	Comparison	0	1	0	0
	Background	1	2	1	0

Table 45 Rationale sample content and author attitude summary table.

8.1.2.3 Evaluation Statements

The rationales and evaluative questions were given to participants via on online survey to record the responses. The rationales were ordered in four sets with each rationale occupying a different position in the task order in each set. This was to counteract any ordering effects of the rationale presentation.

A list of evaluative statements was developed. All participants were asked to evaluate each of the four rationales. The first question required the participants to state whether they

agreed with the argument presented in the rationale, then to rate how confident they were in this agreement level. Most participants rated themselves as holding a moderate (rating of 4) amount of confidence with their initial impressions of the rationales in terms of agreement. Indicating that the rationales made sense and the arguments were obvious and accessible to a lay person. The level of agreement question was also intended in part, to be indicative of persuasion. However, this is somewhat contentious as the content of the rationales may be emotive, and participants may already hold firm beliefs in this area which may influence their perception. The use of a reasonable number of responses should give a more general view of the rationales to counter any highly skewed responses.

The next six questions listed below comprised of statements pertaining to the quality and overall impression of the rationales.

1. I think this person felt confident about their decision
2. I feel this person directed their argument towards another person
3. The rationale contains a good quality argument
4. The rationale is easy to understand
5. The rationale assesses both sides of the argument
6. The rationale is similar to one I would write for this question

The responses were taken in the form of a 1-5 Likert scale to indicate agreement or disagreement with the statement in light of the rationale presented.

8.2 Findings

The agreement indicated by the participants for the arguments presented in each rationale are summarised in Table 46. The least convincing arguments appear to be Rationale One and Four, with participants indicating the highest level of agreement with the arguments presented in Rationale Two and Three. However, this could be a result of Rationale One and Four containing more one sided arguments, thus if the receiver held a position on the topic for which the arguments are not acceptable then they will be rejected. However, as the arguments in Rationale Two and Three construct wider arguments that incorporate both nature and nurture positions as being correct, they may be more likely to be accepted by the receiver as they do not fully contradict either position.

	Percentage Agreement with Argument
Rationale One	35%
Rationale Two	83%
Rationale Three	96%
Rationale Four	39%

Table 46 Summary of participant agreement with the rationale sample.

Rationales Two and Three represented the most two-sided and elaborated arguments respectively. The authors of Rationale Two and Three rated their confidence level and persuasiveness of the rationales as somewhat higher than the other two.

The responses for each of the evaluation measures are summarised in Table 47. The results indicate some interesting trends in the evaluations. The key differences in evaluative ratings and whether these are significant between the rationales will be discussed, alongside possible explanations for the findings.

	Statement	Rationale 1		Rationale 2		Rationale 3		Rationale 4	
		M	Mode	M	Mode	M	Mode	M	Mode
1	How confident are you about your agreement/disagreement with the rationale? *(1=Not 5=Very)*	3.6	4	4	4	4.1	4	4.1	4
2	I think this person felt confident about their decision. *(1=disagree – 5=agree)*	3.9	4	4.1	4	4.4	5	3.9	4
3	I feel this person directed their argument towards another person. *(1=disagree – 5=agree)*	2.5	2	3.2	3	2.9	2	2.9	3
4	The rationale contains a good quality argument.	2.8	2	3.7	5	4.1	5	2.7	1
5	The rationale is easy to understand. *(1=disagree – 5=agree)*	3.8	4	4.3	5	4	4	3.7	4
6	The rationale assesses both sides of the argument. *(1=disagree – 5=agree)*	1.5	1	4.4	5	4.2	5	3	2
7	The rationale is similar to one I would write for this question. *(1=disagree – 5=agree)*	2.1	1	3.5	4	3.6	4	2.1	1

Table 47 Summary of evaluation measure findings for the rationales sample.

In this section the overall responses to the remaining evaluation measures (2-7) are discussed. A Wilcoxon signed ranks test was performed to assess whether the ratings for the statement agreement differed significantly between any of the Rationales.

8.2.1 Statement 2: I think this person felt confident about their decision.

Rationale Three was rated highest for the author feeling confident about their decision based on the rationale. Rationale Three was rated significantly higher than Rationale One (Z=-2.138, p= .033) and Four (Z = -2.072, p = .038). Rationales One and Four were the lowest rated equally in terms of assumed author confidence. The author of Rationale Three did in fact have the highest confidence rating out of the rationales and interestingly, Rationales One and Four had the lowest author confidence ratings out of the four. It is intriguing to note how the author attitudes appear to be successfully translated via the argument structures to the reader. This is an encouraging basis on which to begin to build a model of rationale style argument to assert how the structures within are in fact measurable reflections of the internal processes and attitudes of the author.

8.2.2 Statement 3: I feel this person directed their argument towards another person.

For this question Rationale Two was seen as the most likely to be directed towards another person, with Rationale One as being the least according to the means. This difference is significant (Z = -2.130, p = .033). The author of Rationale One did in fact state that their rationale was predominantly self-directed. This is reassuring in the sense that the awareness of the intended direction of the rationale does appear to have a bearing on the structures and hence how the rationale is perceived by a reader. If the ratings for perceived direction were not indicative of a change in argument strategy, the reader would not be able to detect this intention from the argument. Rationale Two was rated by the author as highly other directed and contains more Contrast relations than the other rationales. This may be a cue to indicate balance for the reader, and again the intention of the author appears to be detected by the reader in view of the externalised structure.

8.2.3 Statement 4: The rationale contains a good quality argument.

The rationale that was rated most favourably for this measure was Rationale Three, with the lowest rated being Rationale Four (Z = -2.454, p = .001). Rationale Three also received significantly higher ratings than Rationale One (Z =-3.070, p = .002). Rationale Two also rated higher than Rationale One (Z = -2.308, p = .021) and Four (Z= -2.543, p =.011) indicating that Rationale Four was considered the poorest quality argument out of the four, with Rationale Three being considered the highest. The author of Rationale Three rated

themselves as having a high confidence in their decision. This rationale also contained the highest number of Backing elements and Elaborations as well as Concessions. Rationale Four in contrast, contained considerably less supporting elements or evidence of a two sided approach to the argument. Again, is seems the perception of argument quality on the part of the reader is influenced by the presence of a balanced argumentative style. However, an alternative explanation may be the language quality in Rationale Four as opposed to the quality of argument. Rationale Four is not grammatically correct in comparison to the other Rationales and it may be this aspect that is more compelling in a negative respect to a receiver regardless of the argument itself.

8.2.4 Statement 5: The rationale is easy to understand.

Rationale Two was rated the easiest to understand with Rationale Four being considered the least comprehensible. The difference in ratings was significant (Z=-2.266, p = .023). However, all four rationales were considered relatively easy to understand with moderately high agreement indicated for this statement. The ease of understanding appears to mirror the ratings for the perceived quality of the argument, as Rationale Four, the least comprehensive of the rationales, was also considered to be the lowest quality in terms of the Toulmin based quality scheme.

8.2.5 Statement 6: The rationale assesses both sides of the argument.

For this question, the rationale with the highest ratings was Rationale Two, which was rated significantly higher than Rationale One (Z = -4.340, p=.000) and Rationale Four (Z= -3.282, p= .001) which were the lowest rated rationales. Rationale Three was also rated highly for this question. Rationale Two, although this was not the rationale that was considered the best in terms of quality, was considered to offer the most balanced argument. This Rationale contained the highest occurrence of Contrast relations (detected by the automated parsers), Classical Concession relations and an instance of a Counter Claim. Rationale One in comparison contained no instances of Contrast relations in any of the approaches. This is encouraging for the potential use of the automated parsers to detect balanced arguments by way of identifying Contrasts. It appears that the presence of these constructs imparts a sense of balance to the reader, which is readily apparent. Interestingly, Rationale Two was considered to be the most persuasive by its author in comparison to the other rationales. This high persuasiveness score is possibly related to an awareness of the balanced

approach. Additionally this rationale was rated by the author as highly other directed. This direction would inevitably give rise to a strategy focussed on persuading others.

8.2.6 Statement 7: The rationale is similar to one I would write for this question.

For this question Rationales One and Four were rated as equally low, indicating participants did not feel that they would write rationales similar to these. Rationale One was rated as significantly lower than Rationale Two (Z = -3.389, p = .001) and three (Z = -3.356, p = .001). Similar findings were found for Rationale Four which was also rated as significantly lower than Rationale Two (Z= -3.034, p =.002) and Three (Z= -3.098, p = .002).

Rationales Two and Three were clearly considered as the most similar to arguments that the participants thought they would write themselves. Rationale Two was considered the most balanced and Rationale Three was consistently rated as the highest Quality, possibly due to the elaborated argument. Although the lengths of all of the rationales were comparable, Rationale Three contained the highest density of Toulmin elements within the argument, which is perhaps an implicit cue for a reader of argument depth and quality. Rationale Two was possibly desirable in terms of its balanced structure; evidently this type of argument is more plausible to the reader and seen as a sign of argumentative competency and confidence.

8.3 Discussion and Limitations

The study was intended to be exploratory in nature to examine the response of the reader to rationales that vary in terms of rhetorical structure. However, a considerably larger number of rationales would need to be investigated in this way to determine true causal relationships between the structures within them and the reader evaluations.

Additionally, a high number of statistical comparisons were conducted, thus the findings should be considered with caution on this basis in addition to the sample size limitation. In an exploratory investigation such as this it is perhaps too conservative to adjust the error rate and it was considered appropriate to consider any significant results using the standard significance level of .05. Therefore, any attempts to draw firm conclusions are not encouraged on the basis of these findings and the trends are reported to prompt further investigation.

A significant limitation in the reuse of the rationales from the empirical research in chapter 3 and 4 is that the rationales are not fully self contained. The rationales were elicited originally in the presence of supporting data to which an author could refer to during construction. As a result of this many rationales refer to information and direct attention to data that is not visible to a future receiver. For example, Rationale One uses the statement "as shown in the Twin Study" to support a nurture claim. This statement refers to text in the information brief that describes how the nurture explanation for aggression is supported by studies using identical twins. A receiver would not be able to see the full extent of this information as they do not have access to the information brief and therefore, could not fully understand or make an assessment of this evidence. The aim of this study was to examine how the presence of argument structures can influence the perception of an argument, as opposed to specifically examining the validity of evidence and knowledge contained within. Although it would be necessary to examine the aspect of knowledge as a consideration in future work, as the reception of an argument is multifaceted, the exploratory nature of this investigation suggests that there is some scope for further research into how the perception of argument can in part, be influenced by its structure.

Rationales Two and Three scored quite similarly across all of the measures. However some interesting subjective differences can be observed. It seems that although Rationale Three, which does not choose a firm position, was overall the most highly regarded both in terms of perceived quality and desire to affiliate with that style of argument, Rationale Two was still regarded as the most balanced argument. This is interesting in terms of previous attempts to propose ideal argumentative structures, as if counter arguments and balance are considered the most powerful, it seems that this is perhaps not the feature to which a lay person would gravitate towards as being the most appealing, although a reader seems to recognise that these aspects represent balance. A fully extended argument that supports a more one sided view seems to be the most valued such as Rationale Two, which initially acknowledges the plausibility of both sides of the argument, eventually concedes to a one sided position. This echoes the concern of Osborne et al (2004a) who have noted that balanced arguments are often the most difficult to elicit and facilitate in students. The apparent natural tendency to prefer more one sided arguments may offer a glimpse as to why students may be somewhat lacking in balanced approaches when constructing arguments. This is of course speculative at this stage, but it reveals a possible need to examine what types of arguments people find most appealing, and how this can impact upon

their own strategies. It may be if these rationales were presented to expert arguers, that an affiliation for the balanced rationale would be more apparent.

It is also encouraging that the author attitudes in terms of the intended direction of the rationale and the confidence in the decision are translated well to the reader. This is useful when considering developing a rationale style argument model, as it would suggest that it would be viable and accurate to assert that assumptions can be made about an author attitude based on the argumentative structures.

The results are also useful in justifying the use of the automated parsers to effectively demonstrate the level of argument balance, especially if Contrasts in particular are considered as a balancing strategy for an argument. The parsers did in fact detect most of these in Rationale Two which was considered by the participants as the most balanced. In addition the author considered this rationale as very persuasive. This suggests that Contrasts are important in the balance and hence plausibility of an argument. The ability of the automated approaches, particularly the PDTB parser to detect these structures bodes well for the development of a semi-automated approach which could help in determining argument quality. This approach may also assist in making assumptions about the author attitude and potential persuasiveness of the rationale based on the frequency of Contrast relations, in conjunction with other important elements. The emerging importance of the Contrast relation will be discussed in the next chapter.

Part Four: Model and Framework Development

This part of the thesis aims to draw together the findings from the experimental work described in part three. This part comprises three chapters. The first, will describe a broad model to assist in visualising the findings from the argument analyses. All three chapters utilise data from the combined rationale corpus which is comprised of all rationales gathered during the experimental work. The findings from the full corpus analyses will aim to address the final three research questions.

The second chapter in this part details the full correlational analyses between the elements and relations identified by the Toulmin, RST, automated and quality framework approaches. These analyses inform both the rationale style argument model and the adapted frameworks which offer a semi-automated, fine grained approach to effectively and efficiently analyse rationale style arguments. The frameworks will build upon the broad Toulmin approach that will be enriched further by incorporating the most relevant RST relations which will enable an assessment of argument quality.

9 Utilising the Findings I: The Rationale Style Argument Model

9.1 Introduction

This chapter will address the sixth research question of whether rationale style arguments can be modelled in terms of typical argument structures. Modelling allows predictions to be made about human behaviour. Analytical models of argument analysis deconstruct arguments, based on their particular components and overall structure. Normative models propose attributes that relate to strength and quality. Descriptive models make explanatory claims about how people tend to argue. As this stage it would be too ambitious to propose a fully descriptive model of argument. However, the findings can be incorporated into a largely analytically based model that demonstrates expected structures and how these relate to the quality (in terms of complexity) of rationale style arguments and decision confidence. There is also room for a discussion on how argument expertise, as well as the perception of direction could impose constraints on argument. This chapter will help to conceptualise some of the processes that influence rationale structure and will assist in visualising typical rationale style arguments informed by a Toulmin and RST based approach.

Many of the rhetorical relations which could potentially be identified are in fact absent from the findings. This lends credence to the proposed idea that rationale style arguments are comprised of a predictable and specific set of rhetorical relations. A theoretical discussion of this is also developed, although this is cautionary as RST is fundamentally an analytical and organisational perspective on text and not a direct indication of cognitive processes or argument quality.

The following sections are included for the purposes of completeness and to act as a reference point for the justification behind the rationale style argument model proposed in section 9.3 and the new adapted quality frameworks outlined in the next chapter.

9.2 Core Relations within the Frameworks

The full corpus used in this analysis comprised of both the OD and SD groups and the Expert sample from the previous work. This section will examine the most frequent relations within the rationales and propose that these are ‘core’ elements of rationale style arguments. The purpose of this analysis is to guide the development of frameworks in the

next chapter. In addition these findings will also enable other researchers who wish to adopt an RST style approach to assist in making predictions regarding expected strategies and to narrow the focus of the extensive relations list by highlighting the most common relations.

9.2.1 Classical RST Core relations

In the Classical RST analysis the full list of possible relations that could be identified numbered 32 in total (see Appendix 2). The RST analysis carried out in the previous studies revealed that only 20 of these relations were found to occur within the corpus. Out of these 20, nine of the relations occurred less than ten times throughout the corpus. Subsequently, for the purposes of developing a standard model, these less frequent relations were removed from the overall core relation set.

The 11 most frequently occurring relations revealed in the manual analysis are categorised and summarised in Figure 29. These relations are grouped into Subject matter, Presentational and Multinuclear categories following the Original Mann and Thompson (1988) framework. It is clear from the analysis that not all rationales could be expected to contain all 11 core relations as the arguments varied greatly in terms of length, depth and complexity.

Presentational
Antithesis
Concession
Evidence
Justify
Subject Matter
Condition
Evaluation
Interpretation
Cause
Elaboration
Multinuclear
Conjunction
Contrast

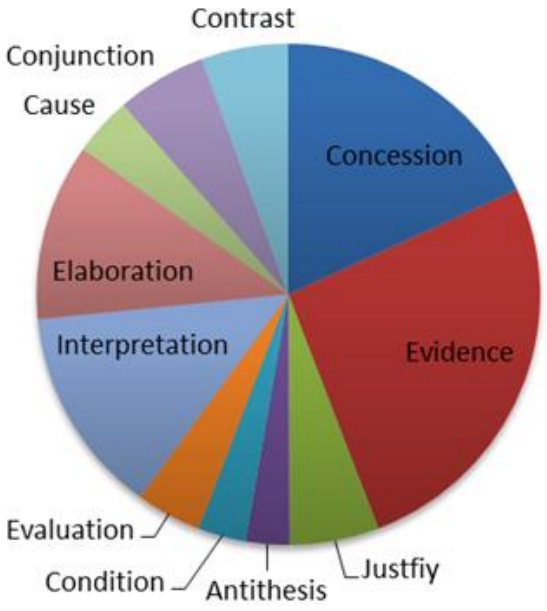

Figure 29 Core Classical RST relations and proportional chart.

The proportional model in Figure 29 shows that rationale style arguments in this context incorporate a wider approach to argument using Concession relations and use supporting elements in the form of Evidence, Elaborations and Interpretations. The less frequent relations appear to act as peripheral constructs which add depth to the argument, adding in causal elements and evaluations of the strength of the data. The multinuclear relations of Contrast and Conjunction do also appear relatively significant to the overall structure. These relations do not have a persuasive nature or an intention to influence the reader's point of view according to the original Mann and Thompson (1988) definitions. Conjunctions are possibly indicative of expansion and extension of information. The use of Contrasts as a strategy to discuss other positions for the argument may indicate that this relation has more of an argumentative property than the original RST framework definition would suggest. A proposed reconsideration of the Contrast relation is detailed in section 10.4.

9.2.2 HILDA Parser Core relations

The findings from the automated HILDA parser indicated that 11 types of relation were present in the corpus out of a possible 18 (full relation list is available in Appendix 3). Again, four of the relations were found to occur less than ten times across the entire corpus and were therefore eliminated from this discussion. The Same unit relation denotes an embedded relation, or a unit of text that relates to a previous tagged text element. As it does not have a rhetorical function associated with it, it will also not be discussed further in this analysis. The resultant six core relations identified by the HILDA parser are summarised in Figure 30.

The Attribution relation is not a rhetorical relation, in that is does not have an argumentative role as such, it simply suggests that a statement has a source attributed to it, either the self, or another agent. As a result of this, the Attribution relation was the second most frequently identified and significantly correlated with many other relations both within other frameworks and within the HILDA findings. This would be a result of the rationale style arguments being constructed on the basis of justifying a personal opinion, resulting in the extensive use of 'I' and the incorporation of evidence to support these views that are often attributed to other research.

Core HILDA Relations
Attribution
Contrast
Condition
Elaboration
Background
Joint

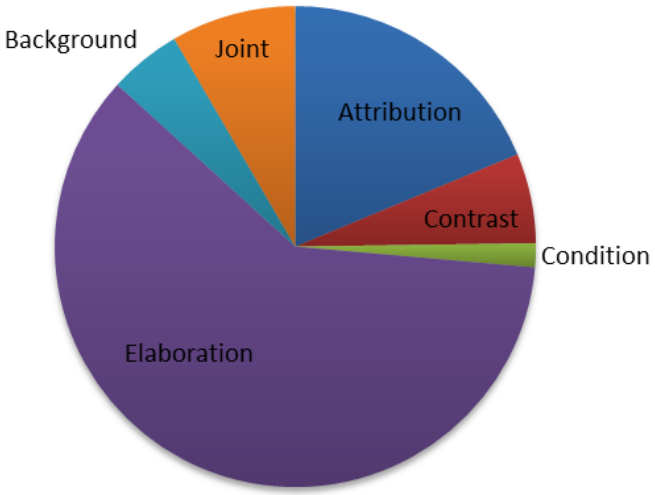

Figure 30 Core HILDA relations and proportional chart.

The proportional chart in Figure 30 shows that the HILDA parser, somewhat overwhelmingly, detects Elaboration relations as the most prevalent relation within these rationale style arguments. Interestingly, the only relation that appears to indicate that the rationales contain a balanced approach, that was reliably detected, is the Contrast relation. The Background and Joint relations are also commonly detected. The Joint relation has no rhetorical nature as such, it pertains to list type elements or the use of 'nor' to signal a Disjunction. This relation is therefore more concerned with the text organisation, rather than the argumentative properties of the rationale. The Background relation offers additional information to enable the reader to understand the main element, it is not causal in nature and it not intended to persuade or indicate importance. The Joint and Background relation appear to have peripheral, expansion properties to the main 'argumentative' nature of the text, but do not, according to the original Mann and Thompson framework, offer additional persuasive functions. However, including additional background information in an argument may have a persuasive impact on a receiver by way of signalling additional supportive information, and thus the level of these structures is still of importance when considering structures which represent quality. In this respect, although Rebuttals are considered a more complex and powerful aspect of argument (Kuhn, 1991), the level of additional information is still important and thus a holistic consideration of a rationale should be conducted.

9.2.3 PDTB Relations Core relations

The findings from the automated PDTB parser indicated that 11 relation types were present in the corpus out of a possible 16. Again, four of the relations were found to occur less than ten times across the entire corpus and as such were disregarded leaving a total of seven core relations that were considered most prevalent in the rationale style arguments.

The text segments labelled as 'Entity relation' by the parser indicated that no separate relations exist, other than the previous relation of which the particular text segment was a part of. This relation is often attributed to a segment of text that follows parentheses. Thus any instances of this were also disregarded from this discussion as they do not signal additional argumentative or rhetorical properties for the overall text. The Entity relations were often identified more frequently in the larger rationales. This may be a result of the increased text size posing a greater difficulty for the automated process. The core relations identified by the PDTB parser in the rationale style arguments are summarised in Figure 31.

Core PDTB Relations
Contrast
Conjunction
Cause
Asynchronous
Synchronous
Condition
Concession

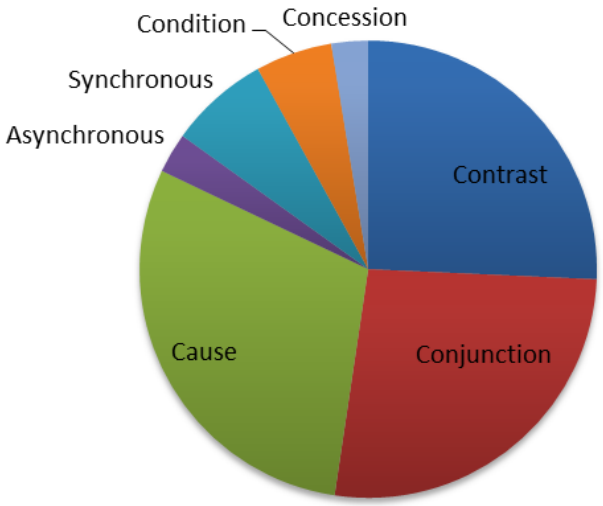

Figure 31 Core PDTB Relations and proportional chart.

The proportional chart in Figure 31 suggests that the balance aspect of a rationale style argument is detected most commonly as a Contrast. A small number of Concessions were identified by the parser, although it appears that Contrast is the most prevalent way of indicating a two-sided argument. The parser also detected various relations that pertained to 'supporting data.' These relations would logically be identified as Conjunction and Cause

as these are the most prevalent relations. As the Conjunction relation is most likely detected by 'and' as a cue in the text, this appears to serve a similar function to an Elaboration relation in the Classical RST approach.

The Cause relation is concerned with contingency, and suggests that one situation is directly influential on another. This relation would be commonly expected in this type of justificatory argument, as the controversial topic requires that causal inferences be drawn between the research material and human behaviour. The original task asked participants to consider the causes of human aggression and decide which is most compelling. The extensive use of Cause relations will be a function of the participants' attempt to demonstrate that their view is the most likely cause attributed to human behaviour. This prevalence may also be expected in such ill-structured and uncertain topics, for which the author needs to draw their own inferences of causality and not rely on absolute supporting data as this type of data is scarce.

The relations of Asynchronous and Synchronous are predominantly identified by the cues 'then' and 'when', it is not therefore surprising that these are also common in these types of arguments as very often temporal style statements are used to provide evidence of actual occurrence. In other words, the author is stating that something is happening or has happened on the basis of the claim. In contrast, the Condition relation suggests that a set of circumstances will need to be in place in order for something to occur and it is not absolute. These types of relations may pertain to a supportive or Evidence type function, whereby they are indicating a causal nature to the data in the argument and thus perhaps making it more compelling to a reader. However, the original RST literature does not suggest that any constraints on the reader occur by using these types of causal or temporal structures.

9.3 The Rationale Style Argument Model

9.3.1 Introduction

This section aims to address the research question of whether rationales can be modelled in terms of expected linguistic and argumentative structure. The discussion will outline a model of typical rationale style argument in an individual context which scopes how the perceived direction effect and the additional consideration of argument expertise can impact on decision confidence and the argument structure in terms of the use of rebuttals. This model is intended to be informative for research into rationale elicitation and support

in terms of suggesting which aspects of the rationale structure need to be supported in order to enhance confidence or the depth of processing of task material.

9.3.2 Model Components

The proposed model for rationale style argument establishes an argument structure, in terms of Toulmin elements, that incorporates the finer grained linguistic structures that may occur, represented by rhetorical structures. The findings from the second study indicate that there were significant structural differences in the frequencies of relations within self and other directed rationales (see section 6.3.8). Therefore the model also addresses the consideration of perceived direction on the strategies used to construct the rationales. Additionally, the factor of argument expertise also features in the model, with suggestions of the impact that expertise may have on rationale structure and quality.

The discussion of the most frequent and typical elements within the rationales in section 9.2 suggests that there is scope for modelling in terms of identifying reoccurring patterns in arguments and a general overview of the way rationales are likely to be constructed.

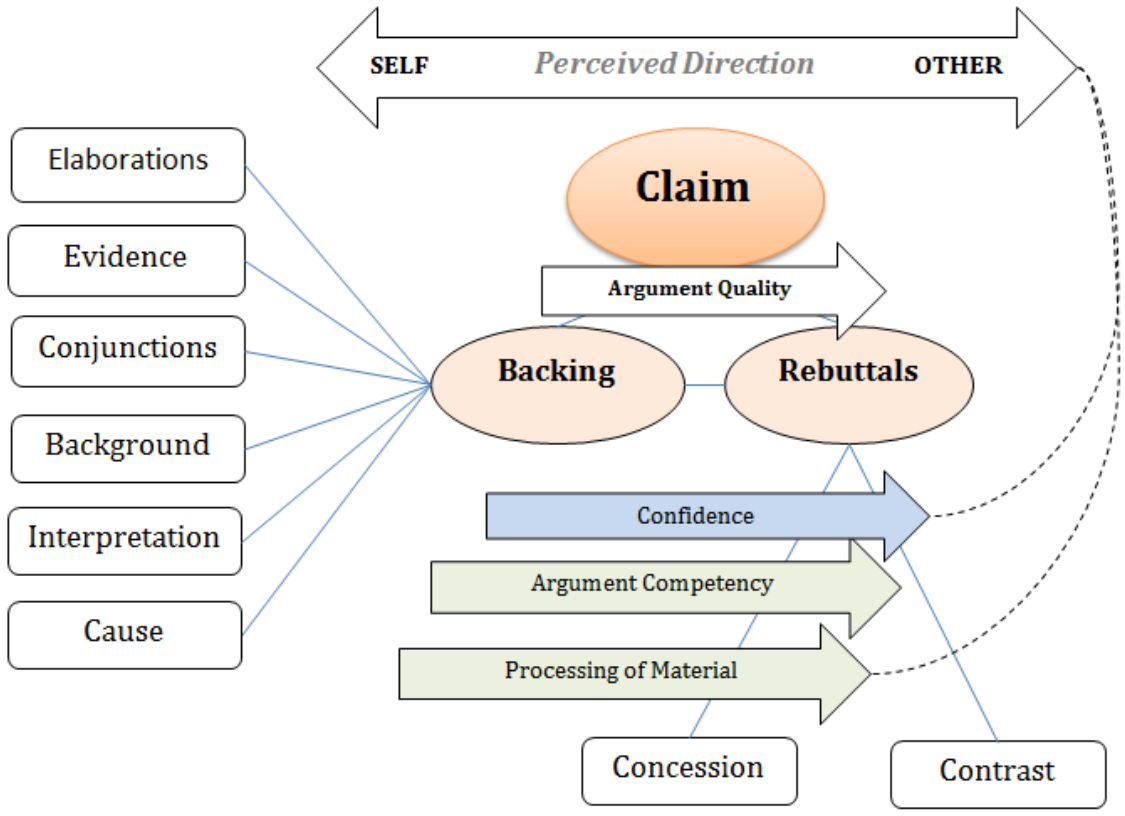

Figure 32 Rationale style argument model.

The model proposed in Figure 32 is comparable to the Kuhn model of argument in that it holds rebuttal as the most powerful aspect in terms of argument quality. In the model the use of the rebuttals is linked to an increase in confidence, depth of processing and the perception of constructing an other directed argument. This theoretical relationship is based on the findings from Chapter 6, that those in the OD group appeared to construct rationales with more Contrast (as detected by the PDTB parser) and Concession relations (detected in the Classical RST analysis) and had significantly higher confidence ratings and raw gain scores. The detection of Contrast in the parsers appears to more reliable than that of Concession, although the definitions (discussed in 6.4.6) would indicate these represent similar linguistic structures, hence both feature in the model as representative of a 'rebuttal' part of an argument.

The model represents associations between argument elements to show how certain elements commonly occur in a typical argument, and the elements that form a balanced rationale style argument. This type of argument encompasses a claim or position, backing in the form of data and phenomena, and some form of critical analysis of the data and its function in terms of supporting or rebutting the main claim or position.

The key features of the model are the contextual considerations of the impact of perceived direction of argument and argument competency. The impact these factors have upon argument quality and attitude towards the decision are outlined. The model proposes a mapping of RST and Toulmin proposed on a superficial level, with a loose mapping of relations onto the Backing and Rebuttal categories. The components and applications for the model are discussed in the following sections.

9.3.2.1 Backing and Rebuttal

This model extends previous attempts to combine RST based approaches with the Toulmin model (e.g. Green, 2010). These attempts did not offer statistical data to support the relationships between the Classical RST relations and the Toulmin elements. This model is based on the findings from the empirical work and an examination of the relationships between the analysis methods detailed in chapter 10.

The inclusion of certain relations in the model was informed by the discussions of the most prevalent relations in section 9.2 and the Toulmin element correlations detailed in Table 51.

The relations included in the model ideally represent those which make up the largest proportions of the argument and thus these are the most likely relations to occur in a typical rationale style argument. The Toulmin element of Backing appeared to correlate with the Evidence, Interpretation and Evaluation relations identified in the Classical RST analysis. As the Evaluation relation did not form a high proportion of the overall rationale (see Figure 29) it was excluded from the model. For the PDTB parser, the relations of Cause, Conjunction, Asynchronous and Synchronous correlated with the Toulmin Backing element. Again, informed by the proportional model in Figure 31, the relations of Asynchronous and Synchronous were excluded due to the lower proportions of these elements in the findings. The HILDA parser relations of Elaboration and Background also correlated with the Toulmin Backing element and therefore also feature in the model.

For the relations that represent the Rebuttal element of the argument, the identification is based on the correlations in Table 51. In the case of the Classical RST relations, Concession correlated with the Rebuttal element. For the PDTB and HILDA parsers, the Contrast relation appeared to be the most significant indicator of Rebuttal elements. The Rebuttal element also correlated with the Attribution and Elaboration relations in the HILDA parser. However, these were not included in the Rebuttal section of the model as the Attribution relation does not have a rhetorical purpose and the Elaboration relation is possibly over identified in the corpus and thus this relationship may be erroneous. See section 11.7.3 for a discussion of parser limitations.

9.3.2.2 Perception of Direction

It appears that the direction of the rationale and the strategies adopted may also lead to deeper and more analytical processing of the material available. The increase in confidence that appears to be function of rationale direction may be a result of this deeper processing, thus inducing a possible learning effect or at the very least, more complex arguments. This awareness that the material has been adequately assessed may influence the acquisition of a positive attitude towards the argument and any decisions made on the basis of it.

The dashed connector between ‘confidence’ and the perceived direction arrow in Figure 32 suggests that confidence increases with a more other directed approach when constructing a rationale. In turn, the second dashed connector indicates that increased processing of material may occur when constructing rationales that are perceived as other directed. The findings in section 6.3.4 support this idea that constructing rationales with a perception of

other may result in deeper processing of material by demonstrating a greater recall of task based information for the other directed group. This effect was not tested in terms of expertise in chapter 7, therefore as this relationship is not established it is absent from the model.

Perceived direction is also linked to argument quality. Quality in this context is measured in terms of the use of rebuttals and this is displayed in the model. The finding from the empirical work that those in the OD group constructed higher quality rationales (section 6.3.7) informs this aspect of the model.

The perception of 'other directed' is linked to a perception of a future use for an argument in the empirical studies due to the wording of the directional prompt. This is worthy of note as there may well be other aspects of future use that could be prompted for to induce an 'other directed' approach that may alter the arguments and attitude. Examples of other types of future use, aside from assisting another in an argument, may be 'to refer to for revision' or 'to publish on a blog.' Even these two types of future use may influence the perception of writing in an other directed style and may alter the approach to a task. Additionally, the future use of 'own reference' could well be a perception held by those in the self directed group and as such these perceptions also need to be measured in order to ascertain the impact they may have on argument and confidence. The actual OD ratings for direction appeared to be around the midpoint of the scale (see section 6.3.1), however the distribution of the extremes in each group did differ (see section 6.3.2) with 43% of participants in the OD group rating their rationales as more other directed compared to only 14% in the SD group. However, as the model stands, the effect of other directed' is more accurately referred to as 'less self directed.' Whether the view of direction can be better prompted or controlled is an area that requires further study.

It is clear that interaction does not always lead to predictable argumentative approaches. Thus it can be expected that the perceived direction may be equally as variable in the influence it exerts on behaviour. The approach to the task is most often determined by the objective and constraints of the task at hand. This assertion is informed by previous work by Xiao (2013) which examined reasoning styles in teamwork. The findings revealed that participants generated more argumentative structures as a task progressed, with more information sharing strategies adopted at the outset. This suggests that the purpose of communication changed over time. This is a consideration that would need further

investigation to ascertain if the goals of the argument context further impact upon the strategies adopted in addition to the perceived direction effect.

9.3.2.3 Argument Competency

It seems that Expert arguers also appear to generate these types of other directed arguments in terms of the higher frequency of Contrast relations (detected by the Parsers) found in comparison to Novice authors (see section 7.3.5). This tendency is represented by the left to right arrow indicating that elements that are related to Rebuttals are more common in expert author arguments. Argument competency is also linked to the quality of argument in the model. This is supported by the finding (see section 7.3.4) that Expert authors constructed higher quality arguments (in terms of the use of rebuttals) than Novices. The direction of the competency arrow also implies that Novice arguers would perhaps be more likely to use Backing elements rather than Rebuttals in their arguments, and would therefore be of lower quality in this respect. This is supported by the finding in section 7.3.4.1, that Novice arguments contained more Backing Toulmin elements than the Expert group.

The confidence aspect of the model is not related to argument competency as the findings in section 7.3.2 revealed no decision confidence effect for expertise.

9.3.3 Applications

The model can be used to analyse important components of the rationale. For example, if a rationale contains a Contrast, this relation indicates a rebuttal which is an indicator of a good quality argument. The associations in the model may enable assumptions regarding author attitude or competency based on the prevalence of particular relations. For example, a novice arguer that is less competent at constructing arguments may be less likely to incorporate rebuttals. In this case rebuttals would need to be explicitly prompted and supported. Similarly, those authors who construct arguments consistently with more rebuttal elements could be considered as having higher expertise in terms of argument skill.

The model could also inform procedures for supporting information recall, as the component relationships suggest that eliciting rationales in an other directed context, that contain rebuttals, could enable deeper processing of material. If the goal of the activity is to facilitate an increase in decision confidence, again the construction of rebuttals would need

to be supported. This is an assumption based on the findings that the OD group constructed arguments with more Contrast relations (an indicator of rebuttal - see Table 52), and this group had higher reported confidence levels. Similarly, if the goal is to elicit balanced, good quality rationales then an other directed context should be established and the construction of rebuttals supported.

9.4 Discussion

The model proposed in Figure 32 is comparable to the Kuhn model of argument in that it holds rebuttal as the most powerful strategy in terms of facilitating the depth of processing and the authors attitude towards an argument. In a similar vein to the Toulmin and Rhetorical approaches the model offers an analytical approach to examining arguments. The model offers additional utility beyond these frameworks as it suggests a potential for supporting strategies that may facilitate better argument quality, such as the use of rebuttals and establishing an other directed approach.

Educators, manufacturers and retailers in particular have an interest in examining human argumentative and reasoning behaviour. It would be useful to be able to make assumptions about an author's attitude based on the features of their externalised argument in a systematic way. This type of approach would be particularly useful when combined with sentiment analysis to examine online reviews. As blogging is fast becoming an indispensable marketing tool, there is a need for a deeper understanding of the rationales that consumers provide for their like or dislike of a product. In this regard, it would be beneficial to assess the strength of the attitudes behind the arguments, and whether these arguments would be potentially influential to other consumers. The evidence throughout the thesis supports the potential use of the rationale style argument model to map out the expected argument scope and thus allow assumptions to be made based on these contextual features of perceived direction and expertise.

The correlational data (see Table 52) that informs the Backing and Rebuttal part of the rationale style argument model will be weaker in some areas, as only one analyst undertook the full Classical RST and Toulmin analysis and there were a large number of comparisons conducted. The most effective way to overcome the accuracy issue would be to have a team of analysts, or a number of consecutive analysts to obtain the majority view or average

agreement. However the findings of the Classical RST do correlate significantly with the automated counterparts where equivalents are present.

The apparent mapping of many of the rhetorical relations from the RST and automated approaches with the Toulmin elements indicated that the development of a framework of argument quality analysis that extends and adapts the Toulmin based quality scheme would be appropriate in order to provide a finer grained, more informative approach. These frameworks will be detailed in chapter 11.

10 Utilising the Findings II: Part One – Analysis Tool Comparison

10.1 Introduction

In order to begin to develop a useful framework for analysing the structural quality of rationales it was necessary to examine how the features of the frameworks utilised in the previous study performed in comparison to one another and which relations positively correlated with the Toulmin based quality analysis scores. This chapter will detail the findings that aim to address the seventh research question of whether Rhetorical Structure Theory (and automated text analysis procedures) can be empirically mapped onto the Toulmin model of argument analysis (see section 10.2.4 for findings). This will add credence to the utility of the automated parsers in conducting a reliable quality analysis that is comparable to a human analyst. This will help to inform and support the production of new semi-automated approaches.

These findings will inform new frameworks described in chapter 11, to enable argument quality analysis from an analytical perspective that can potentially be semi-automated. The original corpus of 99 rationales was combined with the additional Expert rationale sample (N=18) to offer a larger corpus within which to examine the relationships between the approaches. A spearman correlation analysis was performed due to the non-parametric nature of the data. This was only relevant for a particular set of relations, as all three approaches detected relations that were not present, or were without equivalent relations in the other tools. Therefore only the relations which were present in all three approaches (or where there were arguably equivalent relations) were compared. Due to the high number of variables within the frameworks, correlation coefficients below .300 are excluded from the findings and only those with a significance of $p < .001$ will be reported.

10.2 Correlational Relationships between the Frameworks

10.2.1 Classical RST and HILDA Parser

The comparison of the HILDA parser tool with the Classical RST analysis again demonstrates how the automated tool identifies relations that are not present in the Classical analysis and vice versa. In contrast to the Classical RST approach, the parser identifies 'classes' of relations as opposed to specific individual types. This broad approach does make direct comparison to Classical RST difficult; however some general observations can be made. The

summary of correlational relationships between the relevant components can be seen in Table 48.

HILDA Relations	Classical RST Correlates		
	Relation	**Coefficient**	**Sig.**
HILDA Contrast	Concession	.470	.000
HILDA Elaboration	Evidence	.516	.000
	Elaboration	.450	.000
	Concession	.418	.000
	Cause	.318	.000
	Interpretation	.486	.000
	Condition	.318	.000
	Contrast	.310	.001
HILDA Background	Interpretation	.348	.000
HILDA Attribution	Evidence	.339	.000
	Cause	.371	.000
	Concession	.421	.000
HILDA Condition	Condition	.313	.001
HILDA Joint	Conjunction	.326	.000

Table 48 Correlational relationships between the HILDA Parser and Classical RST relations.

10.2.1.1 Presentational Relations

The HILDA parser did not identify any Concession relations in the corpus. The equivalent relation for Concession in the HILDA parser is the Contrast relation, which seems to correlate significantly with the Classical Concession relations. Both relations appear to be cued in the texts by the use of 'but' as a discourse marker.

The Elaboration relation identified by HILDA also correlates significantly with the Classical Evidence relation. This is not surprising as the purpose of an Evidence relation is to provide information with which to convince the reader of a particular point, this may also be demonstrated by giving additional information about or an example relating to the main claim. Therefore, in terms of structure, Evidence and Elaboration may appear very similar

with the Evidence relation having a stronger 'intentional' dimension of persuading the reader. This intentional dimension is problematic in an automated text analysis.

The frequency of HILDA Background relations also appeared to correlate significantly with the Interpretation relations. Interpretation has intentional properties, as it provides further insight from the author point of view, in this case the correlation between this relation and the Background class may be a case of the Interpretation offering additional information in the same vein as Justify and hence the parser views the Interpretation as a circumstantial offering.

The HILDA Attribution and Classical Evidence relation also correlate significantly. This is a logical relationship as the data within an argument is often signalled by offering a source from which the evidence originates. This attribution of source to the evidence provided in the argument would inevitably give rise to the use of discourse markers such as 'shows' and 'suggests' which would trigger the identification of an Attribution relation. This tendency would be particularly common in this style of argument that draws on external sources to support the claims.

The only relation in the HILDA analysis that seems to be without a manual equivalent in the original Classical RST framework is Comparison. The Comparison Class of relations refers to analogies and preference relations which do not form part of the Classical RST definitions. The Comparison relation would be somewhat difficult to explicitly signal or differentiate from a Contrast relation and this may be why it does not have any correlates or equivalents.

10.2.1.2 Subject Matter Relations

The Classical Elaboration relations correlated significantly with the HILDA Elaboration relation. This is not surprising given the HILDA parser's tendency to over identify the Elaboration relation within a text.

There is also a significant positive correlation between the Classical and HILDA Condition relations. The Condition class relation in the HILDA parser definitions incorporates relations that pertain to 'contingency'. This is possibly why the HILDA Cause relation does not appear to be particularly prevalent as the Cause and Result relations are taken up and labelled by the Condition class. This is in spite of a separate class existing in the parser list for Cause

relations. This is possibly why the Condition relation appears to indicate contingency style structures. Essentially, it is a misleading label and possibly explains why no correlations were found for the Cause class of relations.

10.2.1.3 Multinuclear Relations

The HILDA and Classical analyses did not produce any equivalent correlates for the multinuclear relations such as Joint and Disjunction, with the exception of Contrast which is examined in the presentational section.

In summary, The HILDA parser does appear to represent the presentational aspects of the Classical RST approach fairly well, particularly when these relations most often require subjective analysis to identify. However the Elaboration relation in the HILDA parser appears to be too broad and is possibly over identified in the corpus which may have impacted the findings.

10.2.2 Classical RST and PDTB Argument Parser

It is important to bear in mind that the relations, discourse markers and connectives specified in the PDTB parser are not strictly based on RST. However, the definitions in each framework (available in Appendices 3 and 4) appear to demonstrate equivalents in many of the relations. Unlike the HILDA parser, the PDTB parser does attempt to identify and label relations that are not explicitly cued with a discourse marker. It is of interest to examine how this more recently developed set of argument relations correlates with the Classical RST approach and the HILDA parser performance. The summary of correlational relationships between the PDTB parser and Classical RST can be seen in Table 49.

PDTB Relations	Classical RST Correlates		
	Relation	Coefficient	Sig.
PDTB Contrast	Concession Contrast	.628 .414	.000 .000
PDTB Asynchronous	Evidence	.311	.001
PDTB Synchronous	Cause Antithesis	.307 .346	.001 .000
PDTB Conjunction	Evidence Conjunction	.399 .307	.000 .001
PDTB Cause	Evidence Condition Interpretation	.400 .318 .328	.000 .000 .000
PDTB Alternative	Disjunction Summary	.322 .360	.000 .000
PDTB Condition	Condition	.652	.000

Table 49 Correlational relationships between the PDTB Parser and Classical RST

10.2.2.1 Presentational Relations

When comparing the results of the Classical RST analysis with the PDTB parser output, again it is apparent that many relations in Classical RST do not feature in the automated tool. However, there are some significant results which are worthy of note.

It may be the case that the parser tends to favour labelling Contrasts over Concessions, as these are often difficult to distinguish by a human analyst. This idea is supported by a moderate correlation between the PDTB Contrast and Classical Concession relations.

The Classical Evidence relation correlates significantly with the Asynchronous relation. It would not be unreasonable to assume that statements that discuss past and present occurrences are in fact offering evidence to support a claim in the present. This may be a case of the parser identifying these types of constructs in the absence of having a specifically labelled Evidence relation class. The parser definitions do not offer insight into the persuasive nature of these constructs so it is not clear how this may relate to the Classical

Evidence definition, but it is logical that the Asynchronous relation is signalling additional supporting information for a claim in the form of describing events.

In a similar vein, in the absence of a 'catch all' Evidence relation, the PDTB parser relations of Conjunction and Cause may also be appropriate equivalents to Evidence, serving a similar function. The Classical Evidence relation correlates significantly with the Conjunction and Cause relations. As discussed previously, contingency, particularly using the phrase 'because' which is labelled by the parser as Cause, may offer powerful support for the plausibility of a claim. A Conjunction relation offers additional information for a segment, usually identifiable by the use of 'also.' This type of elaboration on a segment could also be seen as offering supporting evidence, as an elaboration can strengthen a claim by signalling importance.

10.2.2.2 Subject Matter Relations

The identifications of the Condition relation by the Classical analysis and PDTB parser correlate significantly. The PDTB Cause relations also appear to correlate positively with the Classical Condition relations. This is not entirely unfeasible as both relations posit a contingency style relationship between two text spans, although one is of course conditional.

The Classical Cause relation also correlates with the Synchrony relation. Synchrony is often denoted by the use of 'when' as a marker, this is also possibly a common way to signal a causal relationship. The author may be describing the circumstances which would lead to a particular event. This would perhaps be common in these types of abstract arguments that are not describing physical events as such, but philosophical probabilities. The PDTB Cause and Classical Interpretation relations correlated positively. As cause suggests contingency or influence, it would not be improbable that a human annotator may interpret a causal statement as an interpretation or indeed as being in conjunction with an Interpretation relation, particularly in light of the abstract material and concepts that the authors discuss in the rationales.

It may appear on the surface that very few subject matter type relations exist in the PDTB parser approach. However, upon examining the parser class definitions some key differences in the approach are evident. For example, rather than being a purely structural

element (as in the Classical RST) the Conjunction relation is given the role of an 'expansion' relation in the PDTB parser. This Expansion level category of relations in the PDTB parser also includes Conjunction, Instantiation, Restatement, Alternative, Exception and List, all of which offer additional information on a concept or artefact. These relations appear to perform elaboration type functions within the text. For example, Conjunction in this case may serve a similar function to Elaboration in the Classical and HILDA analyses. This is supported by a positive correlational relationship between these elements.

10.2.2.3 Multinuclear Relations

It could be argued that the PDTB Alternative relation could be equivalent to the Disjunction relation. In fact, in the definitions which underpin the parser, the Alternative relation class has subtypes which include 'disjunctive' alternatives. A disjunction is often denoted by a 'nor' marker in the text which essentially offers alternative proposition to the main claim that is also not viable, hence a positive correlation was found.

The Classical and PDTB parser findings both revealed significant correlations for the Conjunction as well as the Contrast relations. The positive finding of reasonable correlations between these equivalent relations appears to be promising for the potential of a semi-automated approach to argument analysis.

10.2.3 PDTB Parser and HILDA Parser

The HILDA and PDTB parsers are based on different argument ontologies, with HILDA based upon RST and the PDTB parser based on the concept of argument predicates and structures. However, both identify similar broad 'classes' of relations as opposed to specific types, with similar relations grouped into categories and these categories being used to annotate the text. With this in mind, again, it would be optimistic to expect substantial correlations, but it is worth examining where similarities may lie for the purposes of future argument analysis using these approaches. The summary of correlational relationships between the two automated parsers can be seen in Table 50.

HILDA Relations	PDTB Parser Correlates		
	Relation	**Coefficient**	**Sig.**
HILDA Contrast	Contrast	.563	.000
HILDA Condition	Condition	.524	.001
HILDA Joint	Conjunction	.432	.000
HILDA Background	Synchronous	.401	.001
HILDA Explanation	Cause	.383	.000
HILDA Attribution	Contrast Cause	.401 .460	.000 .000
HILDA Elaboration	Contrast Conjunction Cause Instantiation Asynchronous Synchronous Condition	.471 .366 .539 .373 .377 .331 .300	.000 .000 .000 .000 .000 .000 .001

Table 50 Correlational relationships between the HILDA Parser and the PDTB Parser

Firstly, both the PDTB and HILDA parsers identified the Contrast and Condition relations. A reasonably positive correlation for these relations is evident.

It could be argued that the HILDA Joint relation is equivalent to a Conjunction relation in the PDTB parser, as both would likely use 'and' as a signal and both sides of the relation are considered to be of equal importance. Similarly, Conjunction is part of the Expansion group of relations in the PDTB parser and therefore a Joint relation is arguably an expansion of information that is relevant for two segments of text. However, the Joint relation in the HILDA parser is stated to signal to absence of rhetorical relation between text elements, yet it does indicate a particular structural arrangement that is, that both segments of text are of equal importance to the segment. This is confirmed by a significant correlational relationship between the HILDA Joint and PDTB Conjunction relation, as well as with the Classical Conjunction relations identified.

The Background relation in the HILDA parser results also appears to correlate significantly with the Synchronous relations identified using the PDTB parser. This could indicate that the aspects being measured are also semantically similar, but the specific definitions provided by the HILDA parser regarding the nature of the Background relation are not particularly clear. The PDTB parser literature states that the Synchronous relations are within the temporal class of relations; most often using 'when' and 'then' which may be indicative of the background to a text segment. Again, the HILDA Attribution relation correlates with many of the relations identified in the PDTB analysis. This will be a result of the linguistic nature of a rationale style argument as a personal opinion piece, which will inevitability contain cues such as 'I' and 'think' which are attribution type features of language and will invariably form part of many EDUs regardless of rhetorical nature. For this relation, the individual correlations do not offer any additional insight here.

Finally, the HILDA Explanation and PDTB Cause relations also positively correlate. There is not a clear conclusion that can be drawn between this relationship other than the PDTB parser tends to label Evidence style relations (part of the HILDA Explanation class) as Cause relations. This tendency was highlighted in the previous section that described a correlation between the PDTB Cause and Classical Evidence relations. As the HILDA Explanation class incorporates Evidence, this is not an unexpected finding. Both the HILDA and PBTD parsers have a known difficulty identifying the Cause relations. This and other limitations are discussed later in the chapter.

10.2.4 Toulmin Analysis and Relationships to Other Frameworks

This section details the comparison of the Toulmin model findings and the Classical RST and automated approaches. This analysis aims to address the research question of whether Rhetorical Structure Theory can be mapped onto the Toulmin model of argument analysis. The findings here reveal how the rhetorical relations could be logically categorised in terms of argumentative purpose. Table 51 summarises the correlations between the Toulmin model elements and the rhetorical relations found across all three approaches.

	Classical RST		PDTB Parser		HILDA Parser	
Toulmin Element	**Relation**	r_s	**Relation**	r_s	**Relation**	r_s
Claims	Interpretation Summary	.355* .347*	Cause Alternative	.433* .308*	Elaboration Attribution	.532* .310*
Backing	Evidence Interpretation Evaluation	.582* .465* .321*	Cause Conjunction Asynchronous Synchronous	.410* .367* .358* 308*	Elaboration Background	.610* .381*
Counter Claims	Concession	.442*	Contrast	.324*	Attribution	.397*
Rebuttal	Concession	.471*	Contrast	.507*	Elaboration Attribution Contrast	.414* .407* .344*

*Table 51 Spearman's rho Equivalent Correlations between Toulmin elements and Classical RST and the Automated approaches (*Sig. at .001 level).*

One of the most striking observations in Table 51 is that the Claim and Backing elements from the Toulmin model do appear to correlate with a wide range of relations. This has implications for the utility of the Toulmin model in argument analysis. There is an inherent difficulty in examining correlation data between broad and fine grained constructs as the broader aspects no doubt account for and incorporate many of the finer grained relations. However for the purposes of categorising the relations in terms of argumentative purpose, as either Backing or Balance, the correlations provide informative data. While the Toulmin model is a useful framework for broadly deconstructing arguments, it appears that the labels of Claim and Backing correspond to a rich variety of more complex argument structures. Thus, assigning these broad labels may be reductive, in that much of the intricate features of the argument may be lost.

Overall it appears that both the automated and Classical RST approaches indicate an adequate holistic view of the balance of an argument, which can be mapped onto a Toulmin style model to assist understanding. The mapping provides a clarification of purpose for the

rhetorical relations and the functions they have within an argument, which may not be immediately apparent upon examining the raw parser or Classical RST output. These findings will be used to structure and inform the new argument quality evaluation frameworks in chapter 11.

10.3 Analysis of Quality Framework Findings

10.3.1 Introduction

This section focusses on how the individual relations within an argument may be an indication of the quality overall. The rationales in the corpus were grouped according to the quality scores assigned and the relationships between the relations and the quality levels are examined. The full combined corpus of rationales was utilised for the analysis and discussion. These analyses will inform the hierarchy of relations within the new quality frameworks. For reference
Table 70 through to Table 72 in Appendix 16 present the means and standard deviations for the relations within the rationales, grouped by quality level, for Classical RST, HILDA and the PDTB parsers. Finally, section 10.4 discusses the importance of the Contrast relation for determining argument quality.

10.3.2 Toulmin Model and Quality Score Relationships

The quality scale utilised in the previous study was based upon the Toulmin model of argument, therefore it was expected that the number of Toulmin elements that were manually identified should correlate positively with the quality scores. The expectation was confirmed as the quality scores correlated significantly with the number of Claims, Counter claims, Backing and Rebuttals (Table 52), with Rebuttals as the most highly correlated. This is the element that increases most markedly within the original quality scheme hierarchy.

Quality Score Correlates			
Toulmin	**Classical RST**	**PDTB Parser**	**HILDA Parser**
Rebuttals .651* Backing .581* Counter Claims .526* Claims .346*	Concession .538* Interpretation .441* Cause .426* Evidence .419* Contrast .387* Condition .372* Summary .337* Elaboration .334*	Contrast .546* Cause .455* Conjunction .393* Instantiation .335* Restatement .306*	Elaboration .694* Attribution .543* Background .390* Joint .322*

*Table 52 Toulmin, Classical RST Framework, PDTB and HILDA parser correlates (Spearman's Rho) with quality scores (*significant at the .001 level).*

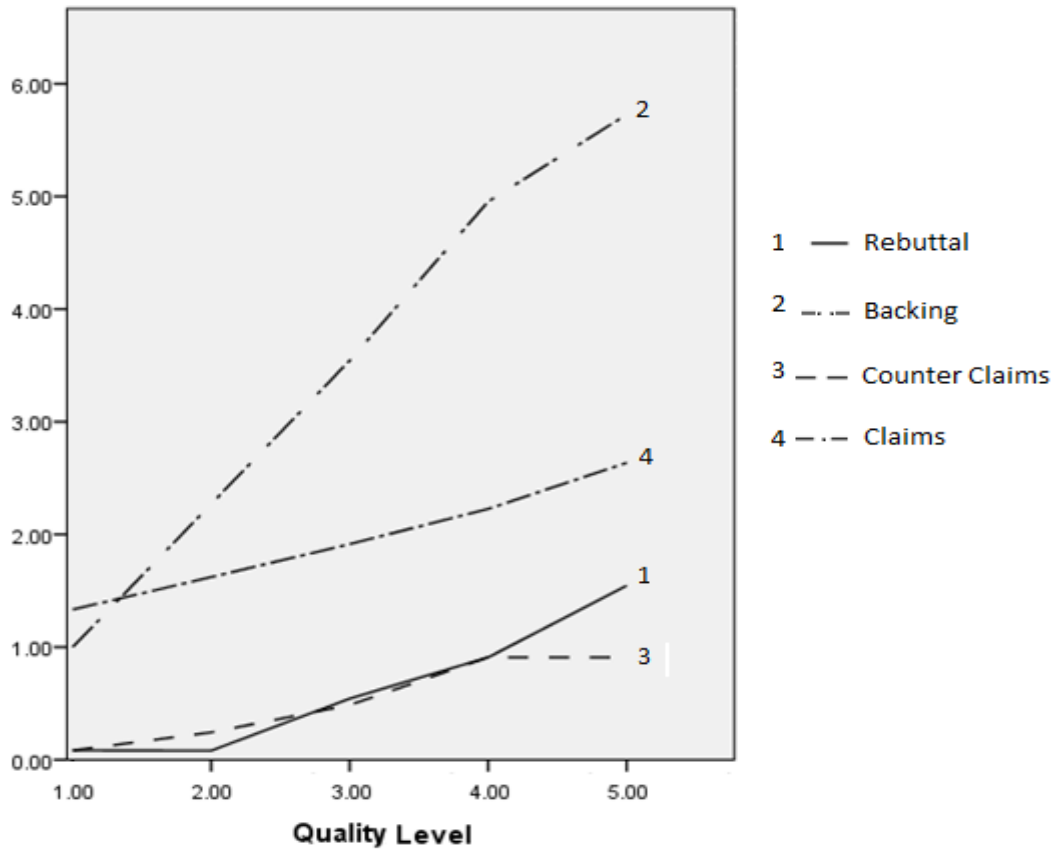

Figure 33 Representation of correlational relationships between the number of Toulmin Elements within each rationale (y-axis) and the quality scores (x-axis).

The Toulmin based quality scores that were assigned to the rationales appear to be validated by the comparisons of Toulmin elements in each level (Figure 33) and these appear to increase fairly consistently with each quality level. This does offer some reassurance for the consistency of the manual Toulmin analysis and quality scheme assessment.

10.3.3 The Classical RST and Quality Score Relationships

The means for the Classical RST relations within each level of quality group are summarised in
Table 70, in Appendix 15.

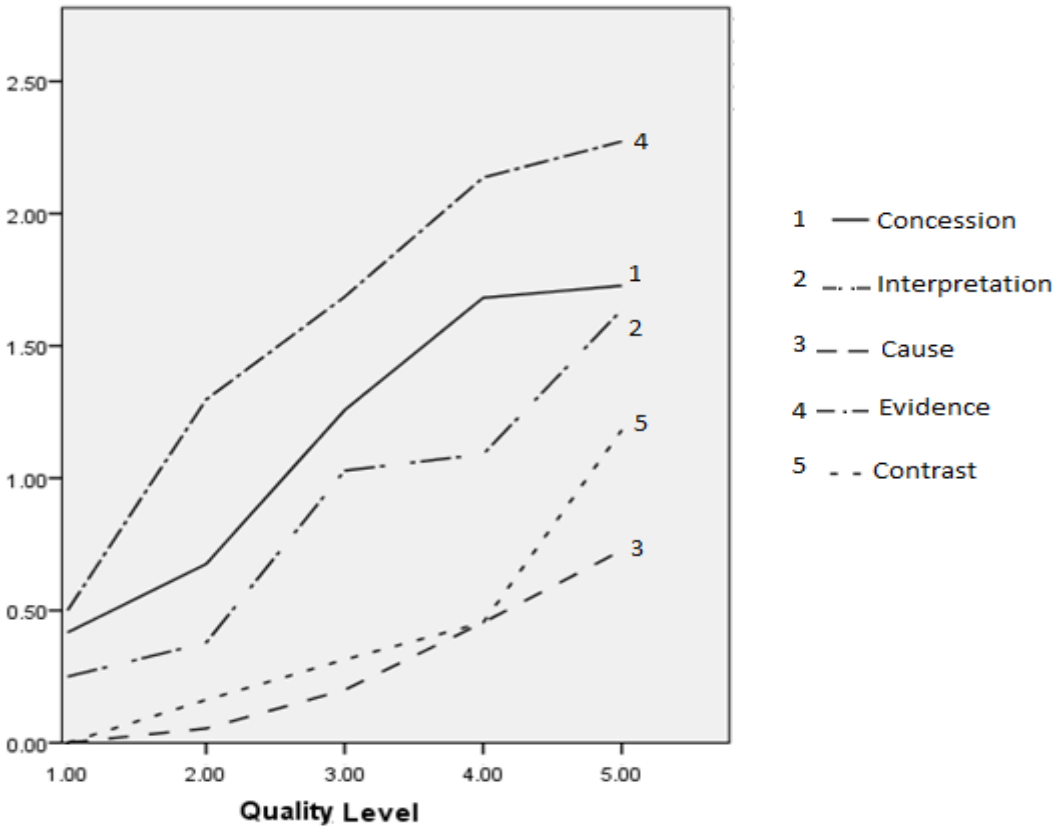

Figure 34 Representation of five highest correlational relationships between the number of Classical RST in each rationale (y-axis) and the quality scores (x-axis).

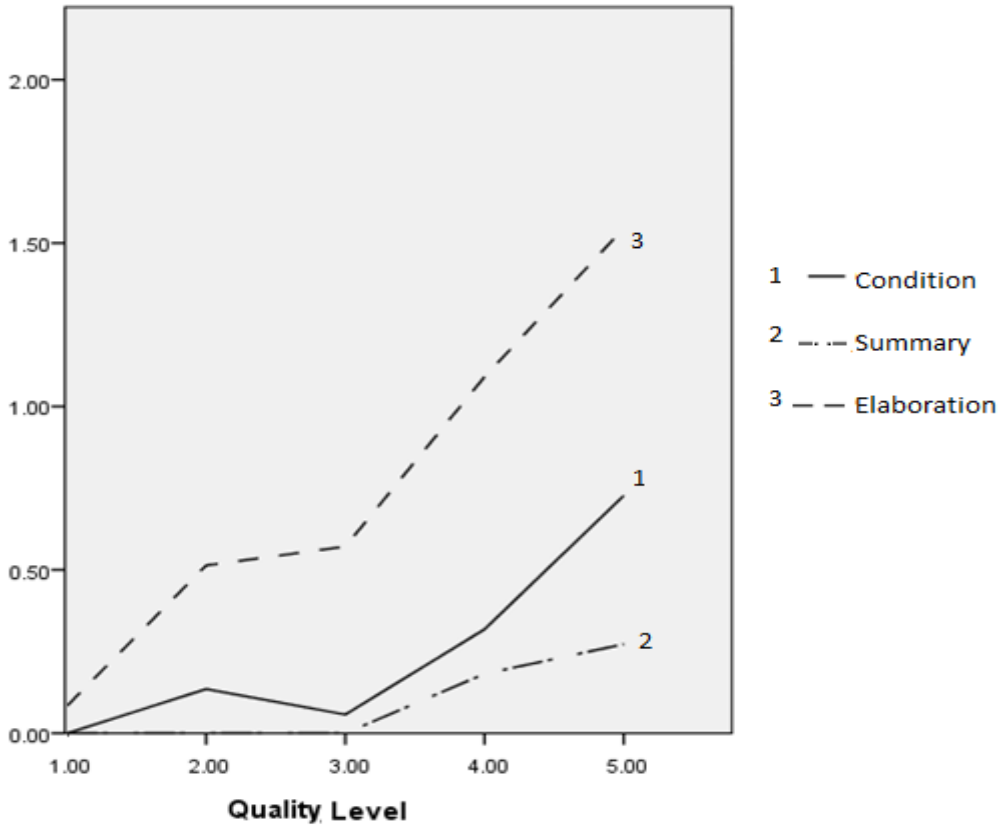

Figure 35 Representation of three lowest correlational relationships between the number of Classical RST relations in each rationale (y-axis) and the quality scores (x-axis).

The quality scores were found to correlate positively with the Concession, Evidence, Elaboration, Summary, Interpretation, Cause, Condition and Contrast relations. Figure 34 and Figure 35 summarise the relationships of the Classical RST relations with the quality level scores. Contrasts and Concessions appear to increase consistently with higher quality arguments, along with Evidence, Elaborations and Interpretations. These elements appear to represent argument Rebuttals and Backing respectively, which mirrors the original quality score hierarchy of balance and depth.

10.3.4 HILDA RST Based Parser

The relations identified by the automated HILDA parser were also correlated with the quality scores. Significant positive correlations (relationships shown in Figure 36) were found between the quality scores and the frequency of Elaboration, Attribution, Background and Joint relations.

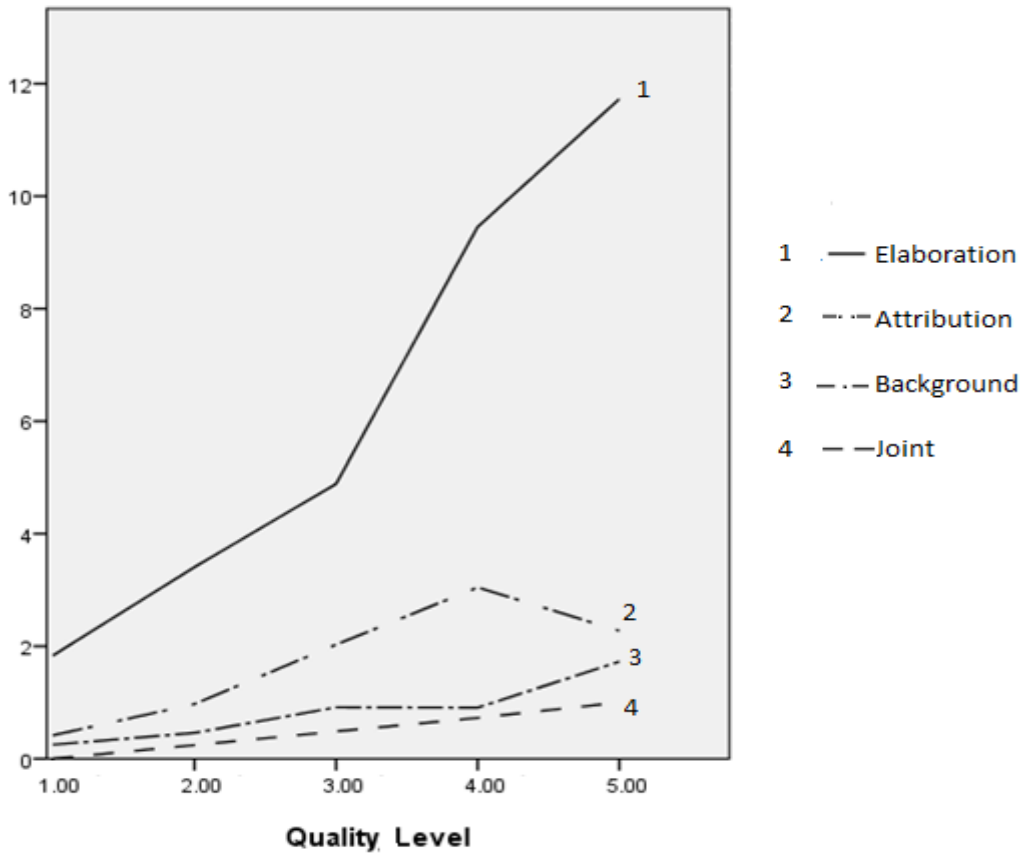

Figure 36 Representation of correlational relationships between the number of HILDA Parser relations from the entire corpus (y-axis) and the Quality scores (x-axis).

The Joint relations are considered as periphery, but are still important elements to offer additional support and discussion for a claim. These elements are not considered to be related to the identification of rebuttals, but could be offered as support or otherwise to a claim presented.

It appears that the HILDA parser may be weaker than the PDTB parser at detecting balanced arguments in terms of identifying rebuttals. Although not reported in Table 52, the HILDA Contrast does correlate with quality score but the relationship is very weak, (r_s (117) = .252, p<.01). As this is the only relation in the parser that appears to be equivalent to a Rebuttal (Contrast correlated with the Rebuttal Toulmin element – see Table 52) it was still considered appropriate to use this relation within the new adapted quality framework as an indicator of the possible presence of a rebuttal type argument within the text.

10.3.5 PDTB Argument Parser

Finally, a number of the relations identified using the PDTB parser also correlated with the quality scores. Significant correlations for the quality scores were found for Contrast, Cause, Conjunction, Instantiation and Restatement relations.

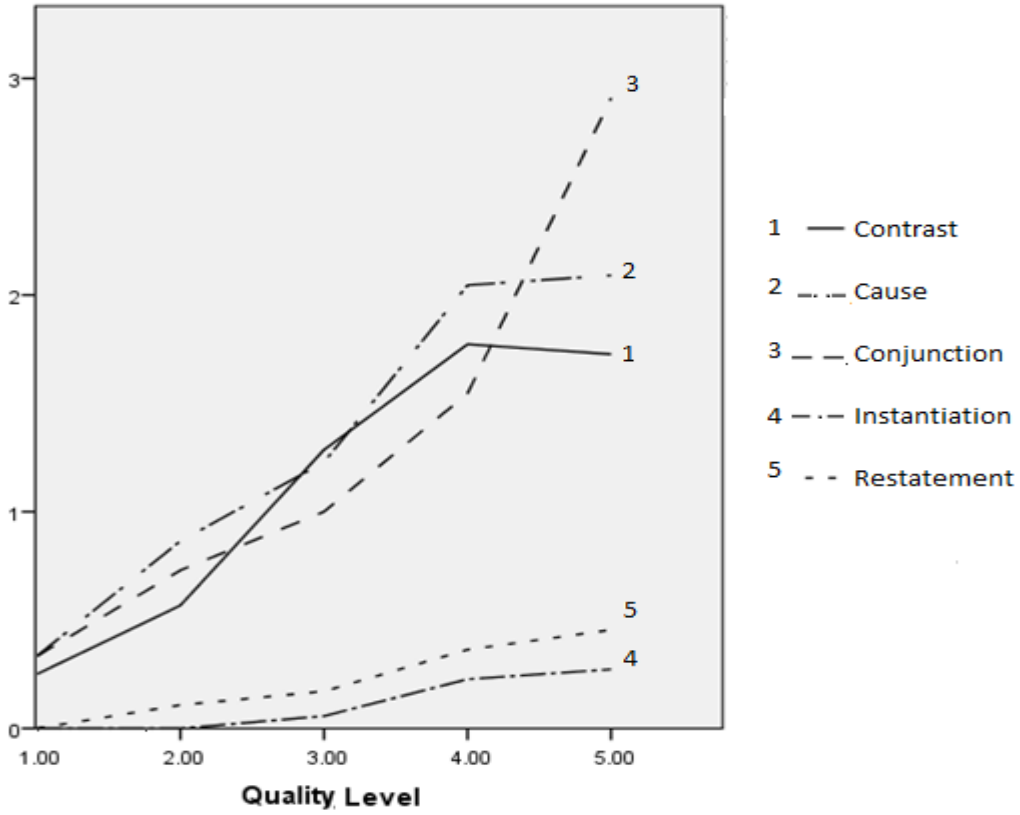

Figure 37 Representation of correlational relationships between the number of PDTB Parser relations from the entire corpus (y-axis) and the quality scores (x-axis).

The distribution of PDTB relations across all quality levels is summarised in Figure 37. It appears that the Contrast relation is related to the Toulmin Rebuttal element (Table 51) and the Classical Concession (Table 49). This suggests that argument balance is detected with the use of this relation. The Concession relation is much less frequently detected by the parser so may not prove an effective method of assessing argument balance from an analyst point of view. The supportive elements of the argument appear to be represented most accurately by the represented by the Conjunction, Cause, and Temporal class of relations as shown by the correlational relationships with the Toulmin Claims and Backing elements in Table 51.

The increase in frequency of these relations in the quality levels suggests more extensively supported and elaborated arguments.

10.4 A Reconsideration of the 'Contrast' Relation

It is worth examining the role of Contrasts at this stage as the research thus far has highlighted some interesting considerations. The Contrast relation forms an integral part of the new quality analysis frameworks and the main reason for this is that it appears to be emerging as an indicator of argument quality (in terms of detecting the presence of rebuttals) which will be discussed further in this section. The Contrast relation as an intentionally persuasive aspect of argument has been potentially underestimated in terms of the Classical RST framework. The evidence for this from the experimental work will be discussed below.

Contrast relations correlate with the Toulmin Rebuttals and argument quality scores (see Table 51 and Table 52) which suggests that this relation is indeed integral to a balanced argument and demonstrates that both sides have been addressed. This conclusion is based on the prevalence of relations detected by the automated parsers in the Expert and the OD groups in comparison to the Novice and SD groups. The findings from the second investigation (see section 6.3.8) demonstrated that those who construct their rationales with the knowledge that the arguments will be viewed and used by others, tended to use significantly more Contrasts within their arguments. Those with a higher level of argumentative expertise also appear to utilise these relations more often than novice arguers. As experts are thought to construct arguments effectively with the assistance of an inner dialogue, like those who construct other directed arguments; it seems that arguments that are produced with a competent strategy and that are intended to influence others are more likely to utilise these relations. This indicates that Contrasts (as detected by the automated parsers) are used in intentionally persuasive, skilled and informative ways. As the quality scores were also significantly different between these groups, it may be that the automated detection of the Contrast relation could potentially be a reliable indicator for quality.

Interestingly, the Contrast relation does not have a presentational type function of persuasion or influence on the reader in the original RST framework. The original RST definition for a Contrast relation is presented in Table 53. As can be seen in the original

definition, the intended effect on the reader of a Contrast relation is primarily to ensure that comparability between two text elements is recognised. The writer is also described as not possessing any internal positive attitude towards any aspect of the contrasting statement, thus the author's intention in using a Contrast is not to increase acceptance or plausibility.

RST Definition of Contrast	Intended Effect on the Reader
No more than two nuclei; the situations in these two nuclei are (a) comprehended as the same in many respects (b) comprehended as differing in a few respects and (c) compared with respect to one or more of these differences	Reader recognizes the comparability and the difference(s) yielded by the comparison is being made

Table 53 Classical RST 'Contrast' definition (Mann & Thompson, 1988)

The Contrast relation definition does not assert any notion of persuasive impact upon the reader to favour one argument over another on the basis of the contrasting segment. Contrasts are often difficult to distinguish from Concession relations. This is undoubtedly due to the similarity in discourse markers that signal their presence. The RST definition for Concession (Mann & Thompson, 1988) is presented in Table 54.

RST Definition of Concession	Intended Effect on the Reader
Writer has positive regard for nucleus. Writer is not claiming that Satellite does not hold; Writer acknowledges a potential or apparent incompatibility between Nucleus and Satellite; recognizing the compatibility between them increases Reader's positive regard for Nucleus.	Reader's positive regard for the nucleus is increased.

Table 54 Classical RST 'Concession' definition taken from Mann & Thompson (1988)

Both relations appear to be drawing attention to two important elements of the text, and either proposing a flaw or violated expectation to the original claim, or contrasting an opposing idea. Although the original definition of Contrast does not state that the Writer or

the Reader will have a positive regard for either of the elements within it, it appears from the findings that they are in fact used to demonstrate that one concept may in fact be more appealing once the contrasting information has been understood. Contrasts serve a purpose in this type of argument context to demonstrate an option has been considered and compared to a favoured option. The Contrast is therefore intended to increase the positive regard for one of these two options and to increase the perception of argument balance which may be a desirable argument trait on the part of the author.

It is also worth noting that the Contrast relation bears a semantic similarity to the Design Space Analysis element of Options. Both structures indicate that two pieces of information have been compared and thus a favourable one highlighted. The findings in the first study demonstrated that Options and Criteria were more common elements in those rationales that were written with prompts that cued possible future interaction with the arguments.

RST Definition of Contrast	**Intended Effect on the Reader**
No more than two nuclei; the situations in these two nuclei are (a) comprehended as the same in many respects (b) comprehended as differing in a few respects and (c) compared with respect to one or more of these differences. *Writer intends the Reader to favour one nuclei over the other. The contrasting intends to strengthen original claim.*	Reader recognizes the comparability and the difference(s) yielded by the comparison is being made. *Reader's positive regard for one of the nuclei is increased.*

Table 55 Redefinition of Contrast in the Classical RST scheme

An extended definition of the Contrast relation, based on the experimental conclusions, is detailed in Table 55. The additions are highlighted in underlined italic fonts. The amended definition would not alter the process of identification of the Contrast relations and this remains intact. Contrasts are still recognised as nuclei that are being compared to highlight similarities or differences. The additions allude to the recognition of the possible intentional aspect of the Contrast relation in terms of establishing a position and explicate the conscious use of the relation to intentionally portray the author's positive regard for one of the nuclei. The additions are considered necessary so the intended effect, or persuasive aspect of a Contrast is not overlooked as is the case with the original definitions. As with the all RST relation definitions this judgement of how plausible it is that a writer intends a reader to

favour one nuclei over another, is still subjective and should be based on multiple analyst judgements. Additionally, it should be noted the Contrast relations as identified by the parsers may not be identical constructs defined in the original RST framework, a concern which would need further comparisons of arguments analysed using the various methods to investigate.

10.4.1 Implications for the Redefining of Contrast

The possible implications for the Contrast relation continuing to be identified as a subject matter relation, with no persuasive properties, include the impact on inferences that are drawn from the use of RST in research. The Contrast relation could be categorised in research as an information sharing style of communication when in fact it may have a more argumentative and persuasive intention. This is particularly pertinent in the use of RST to study collaborative exchanges, these will invariably lead to the production of argumentative structures to some extent and if the Contrast relation is overlooked as being an argumentative construct, the exchanges under scrutiny may appear less argumentative or persuasive than they are intended to be. Thus, to categorise this relation as merely a subject related construct is to underestimate its potentially persuasive and balancing intention in terms of the original RST framework.

The Contrast relation appears in conjunction with the Concession relation in the HILDA class list, this implies that the developers considered the relations as somewhat similar. This is unsurprising as the relations are often difficult to differentiate between for a human analyst and this combining of the relations mitigates this differentiation issue. This possibly increases the accuracy of the parser by reducing the number of relations that exist in the definitions. This issue was also clearly considered by the developers of the PDTB parser, who appear to lean more heavily on the importance of Contrast relations, which are identified more often in a text than Concession relations. This appears contrary to the Classical RST findings.

These findings need to be examined further to tease apart the definitions used by all of the approaches. This warrants further study and manipulation to see if the use of Contrasts is argumentatively powerful in all types of contexts and arguments, and thus whether a redefining is truly warranted, or if its intentional properties are unique to these types of rationale style arguments around a controversial topic.

If Contrasts are integral aspects of good quality or balanced arguments, the production of these relations needs to be explicitly supported and facilitated in argument support systems. Current systems focus on producing pros and cons and counter claims within an argument framework. Contrasts are arguably more sophisticated and difficult to construct than counter claims and may utilise more critical ability, which could in turn increase the depth of processing in a task and thus interaction with the material. If Contrasts are as important as the findings suggest, these need to be explicitly prompted for within these environments.

Previous work has suggested that it is only the presentational categories of relations that hold argumentative power. However the findings discussed highlight the potential argumentative properties of a Multinuclear subject matter relation, contrary to the observations made by Azar (1999), who identified only presentational relations as possessing argumentative properties within examined essay style arguments. It is surprising that the trend of using Contrasts within arguments was overlooked in the work as it likely appeared frequently in these large texts. This type of oversight may be tempting as the presentational relations are labelled as such for their persuasiveness properties for the reader. Thus, the subject matter relations, i.e. non persuasive, could be disregarded as having a more complex argumentative purpose in terms of influence on the reader or intention of the author.

11 Utilising the Findings II: Part Two – Adapted Quality Frameworks

11.1 Introduction

This chapter outlines a set of new quality frameworks that address the final research question of whether RST and automated text analysis tools can be adapted to inform frameworks that assist in the evaluation of argument quality (in terms of the use of rebuttals).

The following discussion introduces new hybrid quality frameworks that incorporate some of basic structures of the descriptive Toulmin model with the fine grained analytical properties of the RST and the text based parsers. The use of RST enables a finer grained approach in combination with the more broad Toulmin style analysis. This should inform a more detailed understanding of the structures within the arguments. Together they offer a more normative approach to assist in determining argument quality, based on the assumption that arguments with rebuttals are of higher quality than those without, thus avoiding circular arguments and reflecting a wider consideration of the evidence and alternatives (Kuhn, 1991). The aim is for an analyst to be able to examine the output from the parser and using the suggested occurrences for the relations, decide upon the best fit level of argument quality. The frameworks will allow for the quick categorisation of arguments and comparison of good versus poor quality arguments depending on the purpose of the analysis.

Currently the RST framework and the Toulmin models have little or no indication of argument power or quality, or how the relations pertain to argument complexity, which would assist in developing a framework of quality analysis. Many researchers adapt the elements and propose their own frameworks based on intuitive and semantic properties of the elements. This analysis will help to reveal which RST relations are more common in complex and higher quality (as determined by the Toulmin based scheme) arguments.

One of the main issues with using Classical RST to analyse arguments is that it is laborious, time consuming and requires training to apply effectively and consistently. It is also very subjective in nature, in spite of the linguistic basis of the framework; it does not routinely specify that certain discourse markers will always denote a particular relation as this is very

context dependent, and many of these markers are implicit. Therefore a plausibility judgment is the main determinant in the RST analytical process.

Two of the frameworks use the text based parsers to analyse quality in a semi-automated fashion. A semi-automated approach would supplement the use of RST in argument analysis by increasing the reliability of the process. The automated parsers are able to label and identify the relations present, and subsequently a framework could be used that outlines how the output relations are an indicator of argument quality. The semi-automated approach to argument analysis would mitigate the time pressure issue of traditional RST methods and will involve considerably less work and knowledge acquisition on the part of the analyst to use. This would ensure that only one plausibility judgement would need to be made, that of a 'best fit' judgement of the output against a quality framework. An analyst would not need to be too familiar with the fundamental theoretical understanding of RST to use the semi-automated approaches and would still be able to be aware of the overall argument quality based on the framework. This could be beneficial in several contexts.

11.2 Framework Structure

11.2.1 The 'Balance and Backing' Dimensions

The new frameworks are split into two dimensions of Backing and Balance for ease of reference and to give a Toulmin based context to the levels and the relations that are identified by the parsers. To an analyst who is not familiar with RST or the automated parser and its argument relations, the actual functions of the relations in terms of argument or author intention are not clear. Thus the framework attempts to loosely categorise the relations based on whether the function is primarily to add 'backing' in the form of supportive data to the argument or whether the function is to demonstrate a two sided argument has been considered, suggesting 'balance.'
These dimensions were informed by the findings detailed in Table 51, of the correlations between the rhetorical relations and the Toulmin argument elements. 'Backing' refers to relations that are statistically correlated with the Claims and Backing Toulmin elements and 'Balance' refers to those relations that are statistically correlated with Rebuttals and Counter Claims. For example, for the HILDA Parser framework, the Backing relations appear to be largely represented with the Background, Elaboration, Attribution relations (shown in Table 51). The aspect of Balance appears to be represented by the Contrast relation which

positively correlates with the Toulmin Rebuttal element. These are the two dimensions of an argument that reflect depth, complexity and thus quality.

The critical relation that will be primarily used to represent Balance and therefore, quality, will be the Contrast relation. This relation is represented slightly different in all three analysis approaches. In the PDTB parser definition in Appendix 4, the Contrast and Concession relations are included in a single class labelled 'Comparison.' This suggests that the developers viewed these relations as performing a similar function in the text. However, the HILDA Parser labels Comparison as a separate construct to those of Contrast and Concession. As both Contrast and Concession are linked under the Contrast class in the HILDA parser this suggests that these relations also, all represent similar functions in the text. The findings from the empirical work (see sections 6.2.3.5.5 and 7.3.5.3) have highlighted that the Contrast relation is a variable that differs between groups who have also been found to differ in terms of quality scores.

Other relations, considered more periphery to the argument and therefore not explicitly concerned with balance, may also still be considered as complex in nature and require more thought on the part of the author to incorporate into the argument. These relations would also therefore be more common in the higher quality levels and may also be indicative of greater argumentative skill. As such, these relations are also incorporated into the quality frameworks where appropriate.

11.2.2 Frequency of Relations within Each Quality Level

The correlations between the relations with the Toulmin based quality scheme levels (Table 52) as well as the significant correlations with the original Toulmin elements (Table 51) inform the structural hierarchy of the new semi-automated frameworks.

The relations outlined in each level are intended to be general expectations based on observations from the data and suggest the most likely occurrences of relations within each level and as such, do not prescribe exact frequencies of relations that could be present. The increase in relations and specific types of relation required for each level is ideally proportional to the overall length of the text.

The numbers suggested for each relation in the new frameworks (see Table 57 for example) are based on an assumption that the rationales being analysed are between 20-100 words.

Given this, variations are expected and therefore a judgement of best fit is required, based on the number of separate relations present, to give an indication of complexity and thus quality in this instance. For reference, Table 70 through to Table 71 in Appendix 16 present the means and standard deviations for the relations within the rationales, grouped by quality level, for all three approaches. However, due to the variation of the relations within the rationales and the large corpus size (117 individual rationales) the raw means are very low, therefore it would not be accurate or appropriate to use these as a direct indication of the expected frequencies of relations within a rationale at a particular level. Similarly, due to the large presence of zero values in the data, the mode and median are equally uninformative. Therefore, the raw means are used as a general indictor of which relations were more frequent in the rationales as the quality level increased.

As level one arguments are considered very basic in all of the new quality frameworks, the levels of all relations can be expected to be between zero and one. The crucial aspect at this level is that no indicators of a balanced approach are present, more specifically, there will be no rebuttals.

	Level 1-2 % Increase	**Level 2-3 % Increase**	**Level 3-4 % Increase**	**Level 4-5 % Increase**
Classical RST	67	35	178	36
HILDA Parser	74	83	45	66
PDTB Parser	50	21	166	24
Mean % Increase	64	46	130	42

Table 56 Proportional increase in means for relations between each quality level across all analysis approaches.

As the means are difficult to utilise in their current form, they are used to suggest how the relations at each level may increase proportionally. Table 56 represents the average percentage increase for relations between each quality level within each framework. These are calculated from the raw means by determining the percentage increase in the mean for each level compared to the previous level for each relation. The percentage increases for each relation were combined and the average calculated. This 'mean percentage increase' represents the average increase in relation frequency within each quality level for all of the approaches.

The percentage increase in relations from level 4 to 5 does represent an anomaly and this may be due to the lower number of rationales in the level 5 group compared to the other quality levels (only 9 rationales). The graphical illustrations based on the findings (Figure 33 to Figure 37) suggest that the relations do increase in a similar expected fashion in concordance with the increases observed in the lower quality levels. For this reason, a base estimation that the number of each relation increases at each quality level by approximately 50-70% is taken. Additionally, the average percentage increase in the relation frequency appears to leap considerably from levels 3 to four. Again this may be a result of the difference numbers for the rationales (37 rationales in the level three group and 22 rationales in level four).

The expected numbers of relations within each level are approximated with the symbol ≈ and a number contained in parentheses (see Table 57 for example). There is a deliberate intention not to be too prescriptive in the exact numbers or the estimation of increases in relations within the frameworks due to the limitations in interpreting the data and the variations that can exist depending on rationale length. Therefore, the presence of Balance relations is the core indicator of quality as opposed to a specification of the exact Backing relations that may be present. The Backing relations are of course required for a good quality argument in conjunction with a balanced approach. This is reflected in the increase of Backing elements throughout the quality levels.

Of course the recommended number and frequency of relations is an aspect of the frameworks that will require further testing and reiteration in order to become more reliable. Additionally, some of the relations occur less frequently than others (see Appendix 15). For example the occurrence of the Restatement and Cause relations in Classical RST and the Asynchronous and Synchronous relations in the PDTB parser (see Appendix 15, Table 70 and Table 71) are considerably lower than that of Elaboration and Contrast and relations, thus the expected percentage increase in the quality levels of 50-70% is not adhered to as rigorously for these relations.

When using the frameworks it is important to take into account complexity in terms of the diversity of relations. For example, a rationale that has eight Elaboration relations may initially appear to fit into the level four of the quality framework. However, if no other relations are present, this rationale should be considered as low in complexity (in terms of relation types) and therefore will be rated as a level one. If the rationale has one Contrast relation, it would then be considered as a level two, and if other relations are present, such

as a Condition or Cause relation, it may fit better into the level three category and so forth. Similarly, for the PDTB framework, a rationale that has many Conjunction relations or even several Contrast relations, would fit into level one and two respectively if no other relations are found.

11.3 HILDA Based Quality Framework

These frameworks can be used in conjunction with the original Toulmin quality scheme to add a more intuitive aspect of the argument overview and a richer sense of argument purpose. The new HILDA based quality framework proposed here (Table 57) mirrors the hierarchical nature of the original Toulmin based quality scheme. The previous analyses inform the structures of the framework by indicating expected frequencies of relations per quality level (informed by Table 70 in Appendix 15) and the correlational findings (Table 52) suggest if and how these relations increase as the structural quality of the argument improves.

The labels within the frameworks that refer to Backing and Balance as discussed above, are intended to aid the order of consideration of the relations. The Backing element in the Toulmin model did appear to correlate with the Elaboration and Background relations (Table 51). Thus, these were considered indicative of the supportive elements of the argument that could also be used as an indicator of quality (in terms of amount of information presented) based on their presence and approximate number. The Contrast relation is utilised in the framework as although no direct correlation was found for the Quality score, this relation did correlate with the Rebuttal Toulmin element (see Table 51) so it was considered appropriate to use the presence and increase of this relation as an indicator of whether a two sided, balanced argument may be present.

The correlational data in Table 52 indicates that the Background, Joint, Elaboration and Attribution relations should systematically increase with each quality level. Additional relations and those that pertain to the 'analyse' categories are also included within the frameworks if they had been found to be a feature of higher quality arguments (see Table 52).

The relation of Same Unit is excluded from the framework as it denotes either an embedded relation or that the argument element is part of a previously labelled span.

	Dimension	Discourse Relations
Level 1	Backing	Most Likely: Few Attribution relations (≈ 1) and Only ≈ 1-2 Elaboration Relations. Possibly one occurrence of a Joint relation. Will not contain a Concession or Contrast relation.
Level 2	Backing	Will contain: Most Likely: Attribution relations (≈ 2) Most Likely: Elaboration relations (≈ 3-4) Will also contain possibly one or two occurrences of a Joint, or Background relation.
	Balance	Will not contain a Concession or Contrast relation.
Level 3	Backing	Will contain: Most Likely: Moderate number of Attributions (≈ 3) Most Likely: Moderate number of Elaboration relations (≈ 4-5) Should also contain at least 2 other relations from any of the following: Enablement, Joint, Condition, Cause, Background.
	Balance	May contain: 1 Contrast or alternatively, 1 Comparison relation to indicate a weak rebuttal.
Level 4	Backing	Will contain: Most Likely: Moderate number of Attribution (≈ 5) Most Likely: Moderate Elaboration relations (≈ 8) Joint (≈ 1-2) Should contain at least 2 or more of any of the following: Background, Cause, Condition, Explanation
	Balance	Should contain 1 Contrast or 1 Comparison relation.
Level 5	Backing	Most Likely: Many Attribution relations (≈ 6) Most Likely: Many Elaboration relations (≈ 12) Joint Relations (≈ 3) Should also contain at least 2-3 instances of any of the following: Background, Explanation, Condition and/or Cause relations.
	Balance	Must contain at least 2 instances or more of any of the following: Contrast and Comparison (indicates multiple rebuttals)

Table 57 New HILDA Parser based quality framework.

The Backing elements in the framework (Table 57), as discussed previously, form the lowest basic level of argument. The more complex relations become more frequent as the level

increases. The Balance aspects of the argument are predominantly determined by the adoption of Contrast style relations into the argument.

The Balance relations reflect the strategies of rebuttals in the Toulmin model which are central to effective arguments. The expected occurrences in the framework for each relation are somewhat general, as the findings that support these demonstrate that relation frequency does vary to some extent within each quality level. The relation types identified and the frequencies of these taken as a whole for the argument should assist in reaching a judgement of a best fit to determine the quality level.

The framework also suggests that an increase in quality level is closely related to an increase in complex peripheral relations such as Conditions, which involve a wider consideration of the supporting data and materials in terms of their implications. This is arguably a more complex and skilled approach to argument than applying supportive data for a single claim.

Although the approximation of frequencies for the Elaboration relations increases through the levels, as a basic indication of argument complexity, it is predominantly a measure of argument length. This is possibly due to the parser's tendency to over assign this relation, particularly if a relation is not clearly signalled in the text. On this basis it would be erroneous to suggest that particular strategies or skills are demonstrated with a higher frequency of Elaboration relations as this may not be the case. Therefore it should be considered predominately as an indicator of argument length.

11.4 PDTB Based Quality Framework

The PDTB based quality framework, presented in Table 58, is again informed by the previous findings. The means for the relations found within the rationales at each quality level (summarised in Table 70 to Table 72) guide the expectations of relations presence and the correlational findings (Table 52) suggest if and how these relations increase as the structural quality of the argument improves. The relation concerned with the Balance dimension of the argument, such as rebuttals is primarily that of Contrast. The Toulmin Rebuttal and Counter claim elements correlated with the Contrast relation and as this is the most reliably detected of these relations by the parser, it features prominently in the framework to reflect balance. The Backing element of the Toulmin model appeared to correlate with the Conjunction, Cause, Asynchronous and Synchronous relations (as seen in Table 51). Additionally, the Restatement relation is included in the framework in spite of the lack of significant correlation with the Backing element as it correlates with the quality scores (see Table 52).

The correlational data (Table 52) indicated that the number of Cause, Instantiation, Conjunction and Contrast relations should increase with each quality level. Additionally, the Asynchronous, Synchronous relations correlate with the Backing Toulmin element (Table 51) thus are included to denote simple supportive elements or data within an argument. The Entity Relation which denotes that no relation could be found is also included in level one in the framework to account for any argument for which no clear discourse marker is present (either explicitly or implicitly) and is thus indicative of a weak or linguistically incoherent argument.

	Dimension	Discourse Relations
Level 1	Backing	May Contain only 1 relation (no more than 2): Either Conjunction or a Cause relation. May contain 1 Alternative to indication a simple Counter Claim. Will not contain a Concession or Contrast relation.
Level 2	Backing	Should contain at least one (ideally no more than 3) of any of the following : Cause (≈1), Restatement (≈1), Conjunction (≈1), Synchronous (≈1). May contain 1 Alternative to indication a simple Counter Claim. Will not contain a Concession or Contrast relation.
Level 3	Backing Balance	Must contain at least 2/3 instances (ideally not more than 4 individual types in total) of any of the following relations: Asynchronous, Synchronous, Cause, Conjunction (≈2) and/or Restatement. Weak Rebuttal indicated by presence of one of the following: Contrast or Concession or Alternative.
Level 4	Backing Balance	Must also contain at least 4 of the following (ideally no more than 7 individual types of relations): Restatement (≈2),, Condition, Asynchronous (≈2),, Synchronous (≈2), Conjunction (≈2), and/or Cause (≈2), relations Should contain up to 1-2 Contrasts (or Alternative or Concession relations) to indicate Rebuttals
Level 5	Backing Balance	Will contain at least 7+ relations comprised of any of the following: Cause (≈2), Restatement (≈1), and Asynchronous, Synchronous (≈1), and Concession and Condition (≈1). May also contain Conjunction (≈ 4) relations. Must contain one of the following: Contrast (≈2) and/or Alternative (≈1) and/or Concession (≈1) relation.

Table 58 New PDTB Parser based quality framework

The expected frequencies stated in the framework (Table 58) for each relation were informed by the procedure in section 11.2.2. The data does suggest that a range of frequencies of relations are possible within each argument level. However the relation types

identified and frequencies of these taken as a whole for the argument should help to inform a best fit judgement for the quality level.

11.5 Classical RST based Quality Framework

The Classical RST framework presented in Table 59 is proposed primarily to give a sense of organisational hierarchy to the original Classical RST relations. As a full, manual rhetorical analysis would need to be conducted prior to using the framework, it does not offer a faster or less demanding approach to argument analysis. It is included in this chapter in the interest of completeness and to act as a supplementary tool to a Classical RST analysis to enable additional judgements by an analyst with regard to argument quality.

The Backing element of the Toulmin model appeared to correlate with the Evidence, Interpretation and Evaluation relations (as seen in Table 51). The Rebuttal and Counter Claims elements also correlated with the Concession relation and therefore this relation is considered to reflect the Balance dimension.

The correlational data (in Table 52) indicated that the Concession, Evidence, Elaboration, Summary, Interpretation, Cause and Contrast relations should systematically increase with each quality level. The approximate numbers for each relation were informed by the data in Table 70 and the proportional increases in each level informed by Table 56.

	Dimension	Discourse Relations
Level 1	Backing	May Contain only 1 relation. (Most typically: Evidence, Justify, Interpretation or Elaboration) Will not contain a Concession or Contrast relation.
Level 2	Backing	May Contain 2 or more instances of any of the following: Evidence(≈2), Elaboration (≈1), Conjunction, Condition, Interpretation. May contain one Antithesis to indicate a simple Counter Claim. Will not contain a Concession or Contrast relation.
Level 3	Backing Balance	Must contain at least 2 (not more than 3-4) instances of: Evidence (≈2), Elaboration, Conjunction, Condition, Evaluation, Summary or Interpretation relations. Weak Rebuttal may be indicated by one of the following (no more than 1): Contrast Or Concession
Level 4	Backing Balance	Must contain 3 or more: Conjunction, Cause, Elaboration, Summary, Evaluation or Interpretation relations. Must contain at least 1: Contrast, Concession or Antithesis relations.
Level 5	Backing Balance	Must also contain at least 4 or more instances of any or all of the following: Cause, Evidence (≈2), Justify, Elaboration (≈2), Interpretation, Summary, Evaluation, Conjunction and Disjunction. Must contain at least 2: Contrast or Concession or Antithesis Relations.

Table 59 New Classical RST based quality framework.

As with the semi-automated approaches, the Classical RST quality framework also categorises relations in terms of balance and backing. The expected frequencies in the framework for each relation are again somewhat general as a range of frequencies and distributions are possible within each argument level. However the relation types identified and frequencies of these taken as a whole for the argument should help to inform a best fit judgement for the quality level.

The increase in quality level is closely related to an increase in Contrast and Concession relations. These relations are complex and are indicative of a two-sided examination of the supporting data and materials. This is, again, arguably a more complex and skilled approach to argument than proposing a single claim with supporting evidence. Reassuringly, the

findings for the Classical RST and the automated parsers appear to concur in some of the key argument quality indicators. This suggests that a manual analysis approach could be superseded by the automated parsers with some confidence.

11.6 Testing the Adapted Frameworks

11.6.1 Between Framework Agreement

To examine how effective these new frameworks may be, they were used to assess the Expert and reduced Novice corpus of rationales. Initially, an independent rater was given output from the HILDA, PDTB and Classical analysis and the new corresponding frameworks to utilise in assigning a quality score to each set of relations. The rater did not have prior experience with the use of RST or argument analysis. This was intended to demonstrate that the frameworks are straightforward to use and that additional skills or knowledge are not essential. The rater was not given access to the original rationales to avoid intuitive ratings for quality. Rationales that are longer in length may automatically be assumed to be of high quality; however this is not always the case. For example, a lengthy rationale may contain many elements of supporting data and elaborations to support a single claim. Whilst this is a good quality argument in respect to supporting one side of an argument, a rationale of equal length that incorporates a discussion of the alternative side of an argument would in fact be considered as better quality in terms of the framework.

	Original	**Classical**	**PDTB**	**HILDA**
Original	1			
Classical	.519 (.561)	1		
PDTB	.662 (.756)	.339 (.409)	1	
HILDA	.627 (.633)	.581 (.582)	.440 (.544)	1

Table 60 Kappa Coefficient Summary for the New Adapted Classical, HILDA and PDTB Frameworks

The new Quality scores assigned by the rater to the rationales, using the Classical, PDTB and HILDA quality frameworks were compared with the Original Toulmin quality scheme scores.

A Kappa coefficient test was conducted to ascertain agreement between the approaches, and the results summarised in Table 60. The unweighted Kappa coefficients are reported with the Cichetti Linear weights reported in parentheses. The weighed Kappa coefficients were included as total agreement using the frameworks would not be expected. Firstly, they required different approaches and used varied strategies and judgement to apply. Secondly, they are carried out by separate raters. Thirdly, the original quality scores were also assigned by a human analyst, thus the original scores may be subject to judgement errors. Finally, the levels of argument quality are not entirely discrete categories and the differentiation between levels one and two or four and five is not necessarily so distinct.

As can be seen in Table 60, the PDTB and HILDA frameworks produced comparable results to the original assessment. The new PDTB and HILDA frameworks both provided a good agreement with the original framework scores. The Classical approach understandably concurred the least with the original quality scores. This is possibly due to the extensive use of subjective analysis in both the relation identification and framework application. However, the automated framework agreement levels are encouraging for the utility of both approaches for assessing argument quality. HILDA offers an RST based approach, which is a well established theory of language coherence. This may be a preferable approach depending on the research goals and a desire for the presence of a theoretical grounding.

Disagreement between the original Quality scores and the automated frameworks may occur due to a misidentification of relations by the parser (or indeed the human analyst). These frameworks do of course rely on a subjective, best fit decision based on the typical frequencies stated, and indeed the output of the parser is not infallible. The agreement levels between the new parsers based frameworks and the original Toulmin based framework appear to be comfortably above chance. This would suggest that these offer a faster and more efficient alternative to the original manual approach. The new frameworks may require less subjective judgements on the part of the rater, as the elements within the argument do not need to be identified, but are indicated by an automated process. This semi-automated approach reduces the overall use of subjective inference and judgement in comparison to the original Quality framework, or a full Classical RST analysis.

11.6.2 Inter-Rater Agreement

To assess how well the frameworks perform when used by independent analysts, a second independent rater was given the same output from the analyses. The decisions for the new

quality ratings for each rationale were compared to those of the first analyst to ascertain the level of agreement using all three new quality frameworks. The agreement for quality scores using all three new frameworks is presented in Table 61.

	Classical Rater 2	**PDTB Rater 2**	**HILDA Rater 2**
Classical Rater 1	.662		
PDTB Rater 1		.694	
HILDA Rater 1			.687

Table 61 Inter rater agreement for the new Classical, PDTB and HILDA quality frameworks.

The agreement between the raters appears to be consistently good using all three new quality frameworks. The variation may be due to the necessity of using subjective judgement to decide on a best fit for each framework level as the guidelines for expected numbers and types of relations are not absolute and thus it may not always be fully clear which level to assign. This type of error or disagreement is to be expected when using inherently variable and human structured text.

11.7 Discussion

11.7.1 Utility of the Frameworks

The frameworks described in this chapter are grounded in the evidence from the research throughout the thesis. The frameworks are founded on the premise that arguments with Rebuttal and Backing elements are of higher quality than those that only consist of claims. This presence of rebuttals indicates a wider consideration of the argument space and prevents circular argument and possibly counter argument by another (Kuhn, 1991). Additionally, this wider consideration of the opposing arguments may help if the goal of the activity is to become familiar with novel material, or to develop argument skills (Yeh & She, 2010).

The semi-automated approaches appear to produce satisfactory agreement with the human judgement analysis using the Toulmin quality scheme. These frameworks offer an alternative approach to a full Classical RST approach. This is desirable for a number of reasons. Firstly, a Manual RST analysis needs multiple raters to be accurate, however the

parsers could potentially act as a second rater in this respect and the findings then evaluated and altered if need be by a human analyst. Additionally, the Classical approach is time consuming and requires training, knowledge and skill to perform effectively. The parsers offer a fine grained approach, in combination with a Toulmin based framework that can be conducted without the need for extensive training if desirable. A human analyst could use the automated parser to deconstruct an argument, but without a framework within which to evaluate the output, a judgement of argument quality would be difficult. These frameworks offer a way to overcome this. RST is also difficult to use for a novice analyst and as a result is not often adopted in research, this is possibly attributable to its origins of being developed for and by linguists. These new frameworks offer a more accessible and approachable method of conducting an RST based argument analysis that would be applicable in a wide range of research domains.

11.7.2 Framework Limitations

The quality frameworks do not of course examine the accuracy of data used, actual persuasive impact of the argument or the relevance and validity of the rebuttals and counter arguments. The quality aspect of concern is indicating whether these elements may be present. To further improve the new frameworks, logic validating measures similar to Walton's schemes could be added as a further subjective knowledge assessment. This is largely dependent on the domain of the argument, and would usually require that an analyst that holds or has access to expert domain knowledge in order to adequately assess the appropriateness of the support offered for a claim or the validity of a rebuttal.

The analyses conducted that inform the framework structure involved a high number of correlational analyses between the relations. The application of corrections in the significance level could be advisable here, however, the high number of comparisons for relations is unavoidable due to the complexity of the analysis methods. Thus a conservative level of correction may result in no significant findings which would arguably be erroneous as there are clearly equivalent relations in the methods which should inevitably correlate. To mitigate this issue, only correlations with a coefficient above .300 were reported, and only if the actual significance was equal to or lower than the .001 level.

The new frameworks are not necessarily intended to be a standalone method for argument analysis. Like many argument analysis frameworks they offer a strategy to evaluate arguments that have the potential to be accessible, modifiable and effective. The actual

argument structures present are considered as providing evidence of critical ability and reasoning styles rather than actual knowledge evaluation as this is a less explored approach. Structural components are an important aspect of argument quality that may form a basis for gaining insight into facilitating critical thinking and supporting the deeper of processing of available resources. Additionally, it must also be acknowledged that the persuasiveness of the arguments is not determined by the quality framework level. This is an aspect that warrants further investigation and analysis as there are various receiver constraints that influence the persuasive impact of an argument and it is an as yet unproven assumption that the presence of rebuttals may be a factor in persuasion.

11.7.3 Parser Limitations

If these parsers are to form the basis of an informative approach to argument analysis it is worthwhile to be aware of the limitations that are inherent within them. There is still some way to go before the automated parsers are as reliable and effective as a manual analysis by multiple raters. A comparison of the analyses available in the Mann & Taboada (2015) website with the relation identification of the HILDA and PDTB parsers shows obvious discrepancy between the labelling of relations. This seems to be most common for the presentational types of relations such as Concession and Antithesis, and also highlights the absence of many of the Classical RST relations from the parsers. This does make direct comparison of findings troublesome. However, the findings in this research do indicate some appropriate correlations between semantically and functionally similar relations. This suggests that these types of parsers, in spite of the limitations could have a useful role to play in expedient argument analysis, which is of particular interest to researchers and also educators who seek to provide timely argument analysis and feedback to learners.

The original proposal paper for the HILDA discourse parser (Hernault, et al., 2010) discussed the performance of the parser in terms of its ability to identify relations across the classes. The performance varied widely, with the Attribution relation being the most accurately identified relation and Cause being the least. This is reflected in the findings for the second study which detected very few of these relations in comparison to the Classical RST, however they did correlate slightly.

The parsers are clearly not infallible in their tagging of the texts. As an example, the PDTB parser inserts connectives where they are implicit in the text which may result in an

incorrect assignment of relation label. However, its bold attempt to identify implicit relations in text is an important step in comprehensive argument analysis especially in view of up to 60% to 70% of naturally-occurring discourse relations being implicit in text (Taboada, 2006). The common pitfall of implicit relations is an issue for any automated parser, as they most often rely on discourse cues indicated explicitly in the text. The HILDA parser appears to default to assigning Elaboration relations in the presence of implicit connectives, this is also reflected in the findings, as the rate of Elaboration relation detection was extremely high in comparison to any other relations.

11.7.4 Conclusions

In summary, like a human analyst, the parsers are not infallible and unlike a human analyst do not exercise subjective judgement to discern the relations. There are obviously pros and cons to both a human and an automated approach. The use of human analysts is of course also fallible, rater and annotator agreement will almost always vary as a result of individual differences, as the impact on the reader is the key aspect for an RST based analysis and thus, the individual perspective will influence the extent and perception of this impact. More pertinently, in terms of use in research, human analysis approaches are time consuming and labour intensive. In order to minimise error a minimum of two analysts to discuss the identification would be desirable. The semi-automated approach to argument analysis would mitigate this time pressure and involves considerably less work and knowledge acquisition on the part of the analyst to use. The analyst would not need to be as familiar with the fundamentals of RST in order to use the semi-automated approaches and would still be able to gain a gist of the argument quality from the framework. These frameworks, particularly the HILDA based strategy should be more appealing to those who are less familiar with Rhetorical Structure Theory, yet require a more accessible but just as informative approach.

The frameworks represent a first step in determining expected argument structures that could pertain to quality in terms of the use of rebuttals. If future testing and refinement of the expected relation numbers within the frameworks is conducted, these can be used algorithmically to potentially fully automate the process of argument quality analysis.

Part Five: Overall Conclusions and Critique

This part of the thesis will reflect upon the experimental work detailed in part three and the development from these findings, of a rationale style argument model and adapted quality assessment frameworks outlined in part four.

This final part will discuss the conclusions that can be drawn from all parts of the thesis and will consider these in light of the research questions that have been addressed systematically through the experimental and conceptual work. The overall limitations of the work will be addressed, followed by a discussion of the wider scope of the findings for understanding human argument and routes for future investigation.

12 Overall Discussion of Findings

12.1 Introduction

This thesis has been concerned with how people construct rationale style arguments and how the perceived direction of these impacts upon the structures within the arguments, task performance and attitude. The research conducted has shown that the perception of direction and of a future use appears to be significant for argument and decision making in the measurable impact on decision confidence and argument structure.

The research has demonstrated evidence for the perceived direction effect impacting on the way information is processed and how arguments are constructed and further, how this process impacts upon confidence. Additional evidence was also found for the level of argument expertise having an impact on argument structure that was comparable to explaining for a perceived other.

The experimental procedures attempted to narrow the focus of the argumentative study to rationale style arguments. The use of a controversial task brief exerted a level of control over the knowledge used and prompted rich arguments. The methodology was designed to reliably control the manipulation of the perception of direction and future use for an argument. The collection of a large corpus of rationales and the extensive argument content analyses conducted has also given rise to new adapted frameworks for argument analysis.

The discussion of the findings and implications will largely be guided by examining the extent to which the original research questions have been investigated.

12.2 Discussion of Research Questions

The **first two research questions** were concerned with exploring an aspect of author perception that was based on the methodological issues in the self explanation literature. The research did not account for how the author's intended direction of the argument, whether for self clarification or for another, may be a critical aspect in determining argument structure and the approach to a task that has been overlooked. The original work examining self explanation used methods that elicited supposedly self directed arguments,

aloud, to an experimenter. This method could easily have resulted in a different type of explanation to one generated in a truly individual context. To examine this, the empirical work in this thesis employed a directional prompt to influence the perception of direction (with the possibility of a future use for an argument) while keeping the actual interaction constant. The findings in the first and second investigations demonstrated that those who had been prompted to write in an other directed approach appeared to construct significantly different arguments, in terms of the use of Rebuttals and Contrasts and were of better quality overall than the self directed group. The actual ratings for direction in the other directed prompt group, did fall at the mid point of the scale indicating that this group could more accurately be described as writing in a 'less self directed' fashion in comparison to the self directed group. In spite of this, the differences in directional ratings was significant between the groups and so indicates that a shift in this perception of direction does appear to alter the approach to argument, both in terms of argumentative strategies used and the perception of decision confidence. This trend is referred to as the perceived direction effect.

The findings from the first investigation were supported by the second study, and demonstrated that those who construct other directed rationales appear to incorporate significantly more Contrast type relations into their arguments. These observations may be indicative of different cognitive strategies that have been reflected in the externalisation of an argument. This would require additional investigation, but it is an arguably more complex process to evaluate and contrast several ideas than to elaborate on a single existing one.

The author's perception of how persuasive they considered their arguments to be was also investigated as an extension to the confidence measure. No significant effects were found for this measure for either perceived direction groups or expertise groups. As persuasion is an aspect of argumentation that is considered as being predominantly ascertained by the receiver, an author's view on the impact of the argument is not a true indicator of persuasive power. It is noteworthy that perceived confidence does appear to be a distinct attitude, unrelated to the sense of whether an argument may be persuasive.

The investigation of the **third research question**, concerned with task information recall, revealed that prompting for other directed arguments increases the complexity of argument structures and that this is triggering a more critical approach to the material and deeper processing. This conclusion is supported by the observed difference in raw learning gain

scores (section 6.3.4). This idea is somewhat hypothetical at this stage, as the differences did not persist when the gain scores were normalised, but it is not unreasonable given the evidence to suggest that this is worth investigating further.

The other directed groups in both investigations also reported higher decision confidence than the self directed groups post rationale construction. The **fourth research question** aimed to investigate whether this increase in confidence was related to the structural features of the rationales. The experimental evidence from the first investigation did suggest that confidence may well be a function of the argument structure (see section 5.3.2.6) in terms of overall length. However, correlations between confidence and rationale length were not observed in the second investigation and no relationships were found between confidence ratings and specific argument structures. This suggests that the confidence effect is possibly related to a perception of a comprehensive argument as a whole, rather than a function of any particular set of structures.

The **fifth research question** of whether the perception of direction could be manipulated successfully, using a written prompt, was investigated via the empirical work in chapters 5 and 6. The empirical work also outlines a novel rationale elicitation task (see section 6.2.3.2 for task procedure) in which the directional prompts were embedded and to enable structural and task variations to be measured based on these. The task successfully elicited rich rationales and also captured the retention of new task material gained as a result of rationale construction. The perception of direction appeared to shift significantly between groups in response to the prompt. This was measured using a Likert response question to assess how well the prompt was attended to and whether the perception of direction had changed. The findings indicated that this was the case (see section 6.2.3.3). A number of issues arose surrounding the wording of the prompt and the inclusion of the wording for 'future use.' These are discussed in more detail in the limitations in section 12.6 and the future work proposals in section 12.5.

A comparison of expert and novice authors was conducted to address **the sixth research question.** The evidence suggested that expert arguers tended to use more rebuttals within their arguments (see Table 37), indicated by the prevalence of Contrast relations. This finding offered an informative comparison with the other directed group in the previous investigation. Similarities were found in the use of Contrasts relations between these groups (detected by the PDTB parser – see sections 6.3.8.3 and 7.3.5.3). This has implications for

describing the cognitive processes that may be responsible for the perceived direction effect, in particular, that explaining in a perceived less self directed manner elicits expert strategies, such as the use of an inner dialogue to aid argument construction. This explanation is theoretical at this stage and is based on the observations of the structural similarities.

The **seventh and eighth research questions** were concerned with a need to address the deficiencies in current argument analysis models. The existing models are largely analytical focussing primarily on structure or are purely domain specific which makes consistent argument research and the drawing of conclusions across findings difficult. It became apparent that there is a need for a step towards developing argument analysis methods that extend the analytical models and offer further insight into the author attitude and perception and suggest how these aspects may influence argumentative strategies.

The Toulmin model, which presents argument elements as purpose based components and RST which presents the rhetorical features of language have been used in tandem to inform the deconstruction of arguments into finer grained constituents (relations) which can be assigned a category of argument purpose (based on Toulmin). The use of the two approaches, both a broad and a fine grained analysis, enabled an extended understanding of both analytical methods. RST relations have been assigned a sense of meaning and hierarchy in terms of argument balance and complexity and the influences these structures have on attitude towards an argument. The Toulmin model has been evaluated with regard to its overly broad categorisation of argument elements. A need for a finer grained approach in conjunction with this method was highlighted. This is primarily due to the Toulmin model's generic treatment of many argumentative structures which can potentially overlook more complex strategies.

The model of rationale style argument that has been developed and discussed in chapter 9, offers a further dimension that is generally absent from current argument models. The model evolved in response to a deeper understanding throughout the thesis and provides an overview of the expected structures within rationale style arguments. The most commonly identified relations are mapped onto a Toulmin structure to offer a broad categorisation of the purpose of these relations in terms of offering either Backing or Rebuttals. The model incorporates the importance of perceived direction and argument competency as features of the situation surrounding argument construction that may have constraints on the types of arguments produced, attitudes and possibly the depth of processing of available materials.

The addition of associations between these factors in the model allows assumptions to be made about an author's attitude based on the structures within an argument (see section 9.3.3 for an example). The findings from the rationale evaluation study in chapter 8, also suggest that explicit argument structure can be an indicator of author attitude that is detectable by the reader. This model offers a more holistic view of argument, extending the previous analytical, dialogue or knowledge based approaches, to include the aforementioned considerations.

The **final research question** was concerned with producing usable and accessible frameworks that enable an analyst without extensive knowledge of RST to conduct a rich analysis of the rhetorical structures present in arguments and to make an informed judgement on the quality of an argument based on these.

The justification for producing frameworks that are used in conjunction with an automated parser is two –fold. Firstly, one could argue that the use of the parsers alone is an adequate method of argument quality evaluation. However, if this was to be the case, the person analysing the parser output would need an in depth knowledge of the nature of the rhetorical relations and how these relate to argument quality. This is most certainly not clear from the current parser output. The new frameworks developed in this thesis will enable the output from these analyses to be assessed which would mitigate the need for in depth knowledge of RST in order to make a judgement, if this was preferred. Both parsers perform reasonably well in comparison to a manual approach. The PDTB parser appears to be more sensitive to text structure given its ability to detect implicit relations. On this basis it could be considered that the framework developed on the basis of this parser was the most successful and informative. However, the HILDA parser is more closely related to the Classical RST approach, and this may be more desirable to a researcher in spite of its tendency to over categorise Elaboration relations.

Secondly, there is potential for these parsers to be utilised to provide argument feedback for users, if used in conjunction with the frameworks to assess argument quality. Currently available argument feedback systems are underdeveloped and do not employ these types of fine grained approaches to argument analysis.

12.3 Additional Findings

The findings from the empirical work also suggested that the role of a Contrast relation may be more powerful within an argument than previously thought. Both the OD and Expert author groups appeared to use these relations more often. The apparent influence that the production of this relation has on the author and the reader of an argument may be more persuasive that the Classical RST definition suggests. An extended definition of the Contrast relation was proposed as a result. This relation was previously considered as a subject matter relation, with no argumentative properties. This finding has implications for future work, particularly if examining arguments or collaborative work as it is common for researchers to group relations into their categories of either presentational or subject matter to assert whether they are intended to be persuasive or offer additional information. Therefore if the Contrast relation continues to be identified as a subject matter relation, it's potentially persuasive intention on the part of the author may be overlooked. However, this finding is incidental in the work and the Contrast relations as identified by the parsers (and highlighted a key difference between the groups) may not be identical to constructs as defined in the original RST framework. This would need further investigation with repeated analyses to examine whether an absolute re-categorisation of this relation is warranted.

12.4 Implications and Applications

The findings and products from the thesis have useful applications, both in further research and in argument support. Some of the potential applications for the work are discussed in this section.

12.4.1 Predicting behaviour

The findings can be used to assist in predicting task performance and attitudes on the basis of argument structure. This can be carried out speedily, with fewer expert analysts than would have previously been required. However the frameworks do not account for receiver involvement in the arguments or where an argument may fall in a receiver's latitude of acceptance when determining argument quality. The frameworks are based on the assumption that balanced arguments have more utility in terms of discussion and consideration of material, but do not make ascertains about how persuasive the arguments may be. This is an aspect which would need further investigation in order for the frameworks to be useful in predicting how persuasive the arguments are.

A popular area of research into public opinion based on internet activity is that of sentiment analysis, to uncover positive or negative affiliations to new ideas or news stories. Research groups that are concerned with ascertaining the public response to environmental issues or energy resources for example, would benefit from being able to identify opinion leaders based on the quality of their posts. These members of the public that respond to news items online, often have an influence on other users' attitudes towards the item, based on their arguments. Argument structures have been shown to be a possible indicator of confidence that is detectable by a reader. The opinion leader arguments may contain certain argumentative strategies which can be identified using the frameworks and the potential influence that these arguments may have, as well as the attitude of the author can be theoretically predicted.

12.4.2 Influencing Behaviour

Argument is also an activity that features in the study of consumer behaviour. The wider scope for the work could be considered to be applicable to the study of recommender systems. The rationale style argument model would suggest the types of arguments that portray a sense of decision confidence to a reader. This is a useful consideration when designing prompts to recommend a product or course of action to a user, which involves a degree of influence and argument.

The frameworks could also be used to ascertain the argumentative competency of a user, and thus offer suggestions for improvement. The argumentative skill of a user needs to be accurately assessed if they are to be supported effectively. The findings also suggest that argument support systems need to ensure that users have a sense of context for their arguments beyond their own use. This may prompt more 'expert' style strategies and thus improve argument skills through practice and the triggering of these improved strategies.

There is scope for using both the model and the frameworks to enhance argumentative skill development and also inform behaviour change prediction in response to good quality arguments, especially if these can be quickly identified and evaluated.

12.4.3 Argument Support

There is potential for RST based tools to be developed that could support argument construction. RST can analyse text based arguments, therefore if a tree formation could be successfully applied to the argument automatically by a tool, a person could view their argument in a similar format to the currently developed graphical representation tools. A RST based tool could allow free text to be entered into the interface, which is a less cognitively demanding activity in itself in comparison to constructing an argument diagram based on predefined nodes. The existing argument support tools almost exclusively force a diagram structure during argument construction which requires additional thought and understanding on the part of the user in conjunction with developing an argument. An RST tool could focus on features which need to be extended and relations that could be added to the free text argument without forcing a predefined structure. The new quality frameworks developed in the thesis, which utilise the parsers, could be used to evaluate the tree diagrams. In order to do this, the rhetorical relations would need to be redefined within the system to enable them to be easily understood by a user, as the current definitions often require deeper linguistic knowledge to fully comprehend. Additionally, the assumption of argument quality being determined by the presence of rebuttals also requires further testing, if it is to be used to enhance argument skill, although this approach has been shown to be effective in previous research (Okumus & Unal, 2012). If indeed, as previous research suggests, a diagram based argument tree approach is beneficial in helping people visualise argument structure, RST based frameworks have something to offer here.

12.4.4 Methodological Implications

The findings discussed in this thesis highlight a potentially crucial methodological flaw in the previous work. In particular, for the work of Chi et al, if self-explaining is a reflective and purely self directed activity it may incorporate difference strategies compared to an other directed activity. However, the work in this thesis has demonstrated that an explanation with no direct interaction but with a perceived sense of future use actually changes the nature of the explanation to incorporate more argumentative elements. This prompts a reconsideration of the previous work on self explanation that suggests that this activity in itself is beneficial, as the benefits may in fact be due to a confounding interaction effect due to the presence of an experimenter which would trigger an argumentative approach.

It is evidently worth bearing in mind how sensitive these perceptions of directing arguments towards the self or others may be and how this can impact upon the quality and structural properties of the arguments produced.

12.5 Limitations

There is a danger within any large body of work of making type I errors; that is declaring a false positive in the findings that are not representative of a genuine effect. Some of the statistical analyses reported a confidence level of .05. However, many findings do report lower values and in addition, the repeated nature of the findings across the studies, such as the variation in the use of Contrast relations between groups, does indicate that the effects observed have credence and these should have bolstered the conclusions to some extent. The findings will of course require additional work to strengthen the effects and uncover insights into the mechanisms responsible. As a first step, a replication of the methodology with multiple analysts would reduce the chance of subjective errors in the analysis. This would allow for the observed tendencies to be more reliably attributed to the influences proposed.

Additionally, there is a large volume of analyses, particularly in the comparison of the structural analysis approaches. There is also a risk of type I errors in this case. To mitigate this, in the analysis method comparison in section 10.2, only statistical findings with a significance level of .001 or lower are reported and additionally, the use of corrections in post-hoc analyses in the earlier work is applied where appropriate. However, in the hypotheses testing, the use of such correction was considered too conservative so as to greatly increase the risk of making type II errors, thus rejecting a true hypothesis. Similarly, in the comparison of the analysis tools, there are a large number of relations within each approach. If a correction were to be applied it would be inevitable that no significant relationships between frameworks inthe analyses would be found. This would almost certainly be an erroneous conclusion as all three approaches use the same rationale data and are based upon similar argument ontologies and thus, some consistency should be expected.

The restricted use of corrections in this case is considered appropriate due to the exploratory nature of the investigations and the use of these analysis approaches which are linguistically based and will inevitably give rise to a large number of variables. Previous

research utilising Classical RST (e.g. Mentis et al, 2009: Xiao, 2013b), usually maintains the integrity of the original framework and does not employ strict exclusion criteria for relations sets in the research in order to remain open minded about the possible variation. The consistent findings of Contrast (and to some extent Concession) relations from the first to the final investigation does support the conclusions that the findings raise - that these relations do vary depending on directional prompting and expertise.

An obvious critique of this work is the use of subjective ratings on the part of the participants to gain insight into the positive regard they hold for their decisions and arguments. This presents an issue of honesty, as it often the case, participants may tend to respond in a way that is considered desirable. To minimise this, participants were not aware of the objectives or nature of the tasks prior to commencing, and no incentives were given at the outset. The consistent findings across the experimental work should offer reassurance that participants were largely transparent regarding their attitudes, but of course the concerns regarding subjectivity will always remain when prompting human participants to describe their inner states.

Additionally, the directional prompting and subsequent measurement for how well this prompt was attended to, may also pose some questions. The directional prompt informed the OD participants that their arguments would provide a future use in assisting others, however this specific perception of future use was not measured in the post decision questions, only the perceived direction. It may be that the perception of future use is the most powerful aspect of the context that influences argument structure and context and not solely whether the argument is self or other directed. As discussed in section 9.3, there are other types of future use that could be prompted and it would be worthwhile to examine and compare how promoting for various types of future use may influence the arguments. Additionally, the SD groups were not prompted for a future use, and there could arguably be relevant uses for an argument generated in self directed manner that were not explored or measured. For the purposes of drawing conclusions from the empirical work, the prompt for future use was originally intended to strengthen the direction effect and did appear to significantly alter the ratings for direction in the OD group into being 'less self directed.' Therefore, the prompt can still be considered as a key factor that has been shown to facilitate the perceived direction effect and result in a shift in perception, though not in an extreme manner.

The frameworks are designed for rationale style argument analysis, to give an indication of quality based on judgement of the automated parser outputs. However, caution would need to be taken, and an analyst would need to be aware that the parsers are not infallible and do not always identify all relations within a piece of text, or label them correctly. In spite of this, the frameworks should still enable a reliable judgement of quality and an overview of potentially important structures within the argument. Additionally, the label of quality is based upon the assumption that arguments with rebuttals are of higher quality than those without (Osbourne, 2004). However, this assumption on which the original quality framework is based, is not yet proven as a domain independent factor, although the linguistic structures within may well be. Arguments with rebuttals may be most desirable in educational contexts or group collaboration whereby wider consideration of resources may be encouraged to ensure better decisions are made and that material is deeply processed. However, in other areas such as consumer behaviour, a two sided argument may not be the most effective when the goal is to influence a purchasing decision. In this respect, argument quality is arguably best defined by the goal of the activity and context in which it is situated, but systematic analysis approaches can still offer ways to identify and differentiate reasoning styles which could be a valid application for a quality framework.

As the determinants of quality for an argument can vary depending on the context, the arrangement of the frameworks is of course open to adaptation. The frameworks used in combination with the rationale style argument model would assist in making assumptions regarding author confidence and the possible depth of processing of the resources available. There is room to combine these approaches with a domain specific knowledge assessment to ascertain the accuracy of the data used. This would add further dimension to the quality assessment.

Finally, although the Toulmin model is well established as a model of argumentation, the use of RST to inform a theory of argument is limited. This is due in part to the linguistic roots of the approach that treats each segment of text as a standalone argument. However, the use of this approach in conjunction with the Toulmin model and the subjective measures have revealed that it does have credence for modelling in terms of identifying reoccurring patterns in arguments and a general overview of the way rationales are likely to be constructed.

12.6 Future work

There are a number of avenues down which the work can be extended. Initially, it would be useful to conduct a direct comparison of these perceived direction groups, who are working alone, with groups who construct arguments in joint and fully interactive contexts. This would provide insight into the potential boundaries of perceived direction and how these perceptions are triggered. It would also enable a rich contrast between the perception of interaction and direct interaction, to see if the strategies for constructing arguments are mirrored.

Similarly, in order to examine the reasons for the OD group ratings being at the midpoint of the scale, it may be useful to examine various prompting approaches to see if the perception of writing for another can be differentiated from writing for a future use as this may have been confounded in the prompts used in this research. Variants of future use such as for revision, publication or analysis could have differing impacts on the structuring of an argument or decision confidence depending on the perception of these. These prompt variations could be tested using the task methodology described in chapter 6, as this appeared to be successful in eliciting rich rationales and is straightforward to implement. Additionally the impact on recall of the varying future use prompts could also being examined using this procedure.

It may also be useful to investigate why novice arguers appear to hold spontaneously varied perceptions of context regardless of the prompt direction, and indeed, at times contrary to the prompt. This tendency is clear from examining Table 24 which show that the mode for the OD and SD group for the directional ratings are identical and both at the midpoint of the scale. In spite of this the SD and OD group differed in their confidence ratings and the use of complex strategies such as Contrast and Concessions. This suggests that a rating given that is contrary to the prompt in the SD group, may not actually represent the same type of perception as a participant rating as other directed in the OD group. It may be, that the inclusion of a clear directional prompt can impact the actual strategies employed regardless of who the author thinks they might be writing for. It would be important to examine when this perception of direction moves from an internal state to actually influencing the strategies adopted in argument externalisation. This could be revealed with a deeper investigation of the arguers internal states, by prompting for in depth subjective descriptions of their perceived argument purpose. It may be that the perception of future use is the most powerful aspect of the context that influences argument structure and not

solely whether the argument is self or other directed. It would be worthwhile to examine and compare how prompting for various types of future use (such as those discussed in section 9.3) may influence the arguments produced and decision confidence.

Further work is needed to determine the nature of the cognitive strategies that lead to measurable differences in argument structure. It may be, for example, that novice arguers adopt the same type of 'inner dialogue' strategy when constructing other directed arguments that expert arguers employ. Although the results indicate similarities in argument structure between Expert and OD groups (such as the increased use of Contrast relations) it is not clear whether this inner dialogue is in fact the strategy adopted by novice arguers. This could be ascertained with think aloud protocols during the argument construction, to ascertain whether a dialogue type strategy is being employed. However, the difficulty with think aloud protocols is the confounding issue of a vocally externalised argument in the presence of others as being 'other directed' and therefore, not the same as a private internal dialogue, which may influence what type of information is given.

One aspect of the tasks in the empirical work which was not examined is that of timing. The precise time taken for the participants to complete their rationales was not measured. In terms of task efficiency it could be argued that an important aspect of good decision making is to complete a task in an optimum time in which the quality of the product, such as an argument is not compromised. As those in the OD group appeared to use arguments with more Contrast and Concession relations, it could be expected that these arguments could take longer to construct as they require examination and assessment of material for both positions. The time taken to complete the task may be a measure of how deeply the task material is processed or, perhaps in a detrimental sense, an increase in cognitive load. It would be important to address the issue of timing in the rationale construction in order to see whether a directional prompt could increase or decrease the time spent on the task and whether this is useful or detrimental to the task outcome.

It would also be of interest to determine why two sided arguments seem to relate to confidence and if they relate to a perception of persuasive power. Again, this could be ascertained by prompting participants to explain the reasoning behind their arguments and which aspects of it they find most compelling. This would go some way to reinforcing the conclusions regarding the rationale style argument model. Indeed, it does appear that in the process of attempting to convince others, we do in fact, convince ourselves.

Finally, the products of the thesis, the semi-automated frameworks need further testing and iteration to validate their utility and efficiency in identifying good quality arguments. The frameworks will need to be tested on rationales from different contexts to ensure that the framework is applicable outside of the domain of a controversial debate. The frameworks are also based upon the assumption that arguments with rebuttals are of better quality than those without. This aspect of the quality framework needs to be tested as this assumption, although supported by research into science argumentation, is largely unproven and would need to be tested thoroughly to ensure the frameworks are fit for the purpose of categorising arguments based on quality. Additionally, the frameworks are intended for use with a limited argument size, around 100 words, so would need to be modified in terms of expected relation frequency if used with much larger arguments.

There is also potential for these automated text parsers to be utilised to provide argument feedback in computer based environment, and to be incorporated into a fully automated argument quality analysis tool, if these assumptions regarding quality can be supported further. Currently, the relation frequency suggestions within the frameworks are not specific enough to be easily applied in an algorithmic sense. Further testing would be needed prior to attempting to fully automate argument analysis based on these.

12.7 Final word

The theory proposed within the thesis is that a perceived other direction for a constructed rationale - if adequately triggered – can cause differences in cognitive strategies. These differences are reflected in structural differences in the externalised argument and reported decision confidence. The perceived direction effect appears to trigger strategies such as the use of Contrasts which were also found more likely to be an expert strategy.

The findings have informed a rationale style argument model to help scope some of the main structural aspects of a rationale and the links found between the perception of direction with confidence, processing of material, argument quality (in terms of the use of rebuttals) and the possible impact of argumentative expertise.

The model could potentially help to predict behaviour, with future iterations and make assumptions about the attitude of an author based on argument features and individual context. The assumptions based on the new model and frameworks would be empirically

based in part. It proposes that the complexity of an argument, which can be influenced by the perceived direction, can also be a possible predictor of confidence, attitude and the depth of processing.

The general findings have confirmed the initial concerns regarding the methodology of the self explanation research. It appears that the self explanation effect is sensitive to the sense of presence or possible future scrutiny of arguments, which may impact confidence, the depth of processing in the task and hence provide additional performance benefits. These benefits are apparent in addition to the established benefits of explaining privately, or rather in a truly 'perceived private' context. This is not to say that people construct poor arguments without a sense of future interaction with the argument, as there are many reasons for constructing a good quality argument in a self directed context, or, that those who intentionally construct arguments with a view to convincing others construct consistently good arguments, but there does appear to be a trend.

The results indicate relationships between the perception of direction and features of the constructed argument. These are summarised in the rationale style argument model and the accompanying frameworks discussed in the previous chapters. A 'generalised contributor' effect of increased confidence may increase task engagement, and thus influence intention, motivation and increased processing within the task. This facilitation effect of argument structure on confidence may also be linked to the concept of critical thinking. In other words, although the normalised gain scores between the self and other directed groups were not statistically significantly different, it does not mean that the effects are insignificant.

Finally, it could be argued that the influential impact of the perception of direction may be considered obvious now it has been explicitly examined. However, it is apparent that this effect has not been considered as a potentially confounding factor in any of the previous investigations into argument quality; either in investigating collaborative work or into the benefits of the self explanation effect. This thesis has demonstrated that perception is a powerful aspect of the activity and would need to be accounted for in future work. The model of rationale style argument provides an overview of the impact that perceived direction may have on argument and attitude.

Thus, it is hoped that the findings assist in future investigations into human argument. The matter of perceived direction needs to be considered as seriously in research as the experimental manipulations that occur in the physical context.

The wider extension of the work may help to support arguments in a variety of contexts offering users an informative evaluation of their work and researchers a rich analysis of human argument. Finally, the thesis proposes a way to introduce accessible RST based argument quality analysis to the field to enable fast, rich and reliable argument quality analysis.

Part Six: Appendices

Appendix 1 Kuhn (1991) Dialogue Based Argument Evaluation Labels

Label	Description
Agree?	A question that asks whether the partner will accept or agree with the speakers claim
Case?	A request for the partner to take a position on a particular case or scenario
Clarify?	A request for the partner to clarify his or her preceding utterance
Justify?	A request for the partner to support his or her preceding claim with evidence or further argument
Meta?	A question regarding the dialogue itself (vs. its content).
Position?	A request for the partner to state his or her position on an issue
Question?	A simple informational question that does not refer back to partner's preceding utterance
Respond?	A request for the partner to react to the speaker's utterance
Add	An extension or elaboration of the partner's preceding utterance
Advance	An extension that advances the partner's preceding utterance
Agree	A statement of agreement with the partner's preceding utterance
Aside	A comment that does not extend or elaborate the partner's preceding utterance
Clarify	A clarification of the speaker's own argument in response to the partner's preceding utterance
Coopt	An assertion that the partner's immediately preceding utterance serves the speaker's opposing argument.
Counter-A	A disagreement with the partner's preceding utterance, accompanied by an alternate argument
Counter-C	A disagreement with the partner's preceding utterance, accompanied by a critique
Disagree	A simple disagreement without further argument or elaboration
Dismiss	An assertion that the partner's immediately preceding utterance is

	irrelevant to the speaker's position.
Interpret	A paraphrase of the partner's preceding utterance with or without further elaboration
Meta	An utterance regarding the dialogue itself
Refuse	An explicit refusal to respond to the partner's preceding question
Substantiate	A utterance offered in support of the partner's preceding utterance
Continue	A continue or elaboration of the speaker's own last utterance that ignores the partner's immediately preceding utterance
Unconnected	An utterance having no apparent connection to the preceding utterance of either partner or speaker

Appendix 2 Classical RST Definitions

Definitions of Presentational Relations			
Relation Name	**Constraints on either S or N individually**	**Constraints on N + S**	**Intention of W**
Antithesis	on N: W has positive regard for N	N and S are in contrast (see the Contrast relation); because of the incompatibility that arises from the contrast, one cannot have positive regard for both of those situations; comprehending S and the incompatibility between the situations increases R's positive regard for N	R's positive regard for N is increased
Background	on N: R won't comprehend N sufficiently before reading text of S	S increases the ability of R to comprehend an element in N	R's ability to comprehend N increases
Concession	on N: W has positive regard for N on S: W is not claiming that S does not hold;	W acknowledges a potential or apparent incompatibility between N and S; recognizing the compatibility between N and S increases R's positive regard for N	R's positive regard for N is increased
Enablement	on N: presents an action by R (including accepting an offer), unrealized with respect to the context of N	R comprehending S increases R's potential ability to perform the action in N	R's potential ability to perform the action in N increases

Relation Name	**Constraints on either S or N individually**	**Constraints on N + S**	**Intention of W**
Evidence	on N: R might not believe N to a degree satisfactory to W on S: R believes S or will find it credible	R's comprehending S increases R's belief of N	R's belief of N is increased
Justify	none	R's comprehending S increases R's readiness to accept W's right to present N	R's readiness to accept W's right to present N is increased
Motivation	on N: N is an action in which R is the actor (including accepting an offer), unrealized with respect to the context of N	Comprehending S increases R's desire to perform action in N	R's desire to perform action in N is increased
Preparation	none	S precedes N in the text; S tends to make R more ready, interested or oriented for reading N	R is more ready, interested or oriented for reading N
Restatement	none	on N + S: S restates N, where S and N are of comparable bulk; N is more central to W's purposes than S is.	R recognizes S as a restatement of N
Summary	on N: N must be more than one unit	S presents a restatement of the content of N, that is shorter in bulk	R recognizes S as a shorter restatement of N

Reproduced from Mann and Taboada (2015) N = Nucleus, S = Satellite, W= Writer, R= Reader

Definitions of Subject Matter Relations			
Relation Name	**Constraints on either S or N individually**	**Constraints on N + S**	**Intention of W**
Circumstance	on S: S is not unrealized	S sets a framework in the subject matter within which R is intended to interpret N	R recognizes that S provides the framework for interpreting N
Condition	on S: S presents a hypothetical, future, or otherwise unrealized situation (relative to the situational context of S)	Realization of N depends on realization of S	R recognizes how the realization of N depends on the realization of S
Elaboration	none Subtypes: • set :: member • abstraction • whole • process • object • generalization	S presents additional detail about the situation or some element of subject matter which is presented in N or inferentially accessible in N in one or more of the ways listed below. In the list, if N presents the first member of any pair, then S includes the second:	R recognizes S as providing additional detail for N. R identifies the element of subject matter for which detail is provided.
Evaluation	none	on N + S: S relates N to degree of W's positive regard toward N.	R recognizes that S assesses N and recognizes the value it assigns
Interpretation	none	on N + S: S relates N to a framework of ideas not involved in N itself and not concerned with W's positive regard	R recognizes that S relates N to a framework of ideas not involved in the knowledge presented in N itself
Means	on N: an activity	S presents a method or instrument which tends to make realization of N more likely	R recognizes that the method or instrument in S tends to make realization of N more likely

Relation Name	**Constraints on either S or N individually**	**Constraints on N + S**	**Intention of W**
Non-volitional Cause	on N: N is not a volitional action	S, by means other than motivating a volitional action, caused N; without the presentation of S, R might not know the particular cause of the situation; a presentation of N is more central than S to W's purposes in putting forth the N-S combination.	R recognizes S as a cause of N
Non-volitional Result	on S: S is not a volitional action	N caused S; presentation of N is more central to W's purposes in putting forth the N-S combination than is the presentation of S.	R recognizes that N could have caused the situation in S
Otherwise	on N: N is an unrealized situation on S: S is an unrealized situation	realization of N prevents realization of S	R recognizes the dependency relation of prevention between the realization of N and the realization of S
Purpose	on N: N is an activity; on S: S is a situation that is unrealized	S is to be realized through the activity in N	R recognizes that the activity in N is initiated in order to realize S

Relation Name	**Constraints on either S or N individually**	**Constraints on N + S**	**Intention of W**
Solutionhood	on S: S presents a problem	N is a solution to the problem presented in S;	R recognizes N as a solution to the problem presented in S
Unconditional	on S: S conceivably could affect the realization of N	N does not depend on S	R recognizes that N does not depend on S
Unless	none	S affects the realization of N; N is realized provided that S is not realized	R recognizes that N is realized provided that S is not realized
Volitional Cause	on N: N is a volitional action or else a situation that could have arisen from a volitional action	S could have caused the agent of the volitional action in N to perform that action; without the presentation of S, R might not regard the action as motivated or know the particular motivation; N is more central to W's purposes in putting forth the N-S combination than S is.	R recognizes S as a cause for the volitional action in N
Volitional Result	on S: S is a volitional action or a situation that could have arisen from a volitional action	N could have caused S; presentation of N is more central to W's purposes than is presentation of S;	R recognizes that N could be a cause for the action or situation in S

Definitions of Multinuclear Relations		
Relation Name	**Constraints on each pair of N**	**Intention of W**
Conjunction	The items are conjoined to form a unit in which each item plays a comparable role	R recognizes that the linked items are conjoined
Contrast	No more than two nuclei; the situations in these two nuclei are (a) comprehended as the same in many respects (b) comprehended as differing in a few respects and (c) compared with respect to one or more of these differences	R recognizes the comparability and the difference(s) yielded by the comparison is being made
Disjunction	An item presents a (not necessarily exclusive) alternative for the other(s)	R recognizes that the linked items are alternatives
Joint	None	none
List	An item comparable to others linked to it by the List relation	R recognizes the comparability of linked items
Multinuclear Restatement	An item is primarily a re-expression of one linked to it; the items are of comparable importance to the purposes of W	R recognizes the re-expression by the linked items
Sequence	There is a succession relationship between the situations in the nuclei	R recognizes the succession relationships among the nuclei.

Appendix 3 HILDA Parser Class Relations List

Class Level (Main Labels)	**Description**	**Sub Levels**
Attribution	Statements that refer to agent origin e.g. "I" "them" "they"	attribution, attribution-negative
Background	Additional information offering a basis for a statement.	background, circumstance
Cause	Elements indicating contingency of one element on another	cause, result, consequence
Comparison	Elements that compare two elements to demonstrate differences/similarities	comparison, preference, analogy, proportion
Condition	An elements is dependent on the circumstances of another being realised.	condition, hypothetical, contingency, otherwise
Contrast	Elements that compare two elements either in a neutral way or to demonstrate a flaw or alternative argument	contrast, concession, antithesis
Elaboration	Elements offers additional information on another – either example or more specific object.	elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition
Enablement	Element that informs how an action can be undertaken.	purpose, enablement
Evaluation	Elements that assess the validity of another either positively or negatively.	evaluation, interpretation, conclusion, comment
Explanation	Element that offers additional	evidence, explanation-

	information in order to strengthen a claim made.	argumentative, reason
Joint	No intention, usually a list of items.	list, disjunction
Manner-Means	Usually indicated by the discourse marker 'by'.	manner, means
Topic-Comment	Element offers a solution to a problem or proposes a question.	problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question
Summary	Element summarises information stated earlier in more concise manner.	summary, restatement
Topic Change	Change of topic within the argument.	topic-shift, topic-drift
Temporal	Time contingent statements, such as 'before', and 'after'	temporal-before, temporal-after, temporal-same-time, sequence, inverted sequence
Same Unit Text Organisation	Indicates that an element is part of a previous relation – no additional purpose is given with this label – often used for parentheses.	Same Units No new Relation

Appendix 4 PDTB Parser Class Relations List

Main Class	Description	Sub Types (Labels)
Temporal	Elements that describe 'when' and 'then' time contingent statements	Synchrony Asynchronous
Contingency	Elements that discuss causality and contingency relationships.	Cause Pragmatic cause Condition Pragmatic Condition
Comparison	Elements that compare two elements to demonstrate differences, flaws or similarities.	Contrast Pragmatic Contrast Concession Pragmatic Concession
Expansion	Elements that offer additional information, summarisation or examples.	Conjunction Instantiation Restatement Alternative Exception List

Appendix 5 Eliciting Self and Other Directed Rationales: Study Brief

You have been put in charge of a New Drug Trial treating patients with a two similar strains of bacterial infection that has proven resistant to previous treatments. You have a choice of 3 pre-trialled drugs to administer to 2 patients who have presented symptoms of infection for over a year. The drugs have been previously trialled for similar bacterial strains and the success rates for the drugs in treating the current infection have been estimated based on this data. Your patients each have a profile of health, detailing severity of symptoms, current health and possible impact of side effects from the treatment. Similarly, the success rates of the drugs, possible side effects including severity and the positive response rate of previous patients to the drug. You should choose the drug that you think will be most beneficial to each patient in the trial based on the information given.

Patient A Profile: Patient A has suffered from symptoms caused by the alpha-bacterium strain and is considered to have acute (severe) physical symptoms including nausea, gastric pain and high blood pressure (hypertension). Their response likelihood to a drug treatment is considered to be good. Their susceptibility to side effects is considered to be high and may affect their quality of life whilst undergoing treatment

Patient B Profile: Patient A suffers symptoms of both the alpha and beta-strains of the bacterium and is considered to have moderate symptoms (average) including mucus production in the sinuses, minor short term headaches and acid reflux. Their response likelihood to a drug treatment is considered to be low. Their susceptibility to side effects is considered to be low and should not affect their quality of life whilst undergoing treatment

Drug 1: Is proposed to be most useful in treat alpha-bacterium infections and is considered to have a success rate of 60% on average. The side effects considered moderate in nature and very common in patients who use the drug. The side effects include nausea, mood swings, and skin sensitivity

Drug 2: Is proposed to be most useful in treat beta-bacterium infections, but could be used in some instances of alpha-bacterial infection and is considered to have a success rate of 78% on average. The side effects considered severe in nature and very rare in patients who use the drug. The side effects include bruising, abnormal blood clotting, difficulty swallowing and inflamed mucus membranes

Drug 3: Is proposed to be most useful in treating both alpha and beta-bacterium infections and is considered to have a success rate of 54% on average. The side effects considered low in nature and very common in patients who use the drug. The side effects include drowsiness, loss of appetite and mild irritation of the bowel.

Please provide your rationale for your decisions here, explaining the criteria you considered as key for your decision, the questions you thought were vital to ask and the options you considered.

Self Prompt: Your rationale will be kept private and not shared.

Other Directed Prompt: Your rationale will be shared with others in the future to assist in decision making.

Appendix 6 Eliciting Self and Other Directed Rationales: Rationale Samples

This drug is considered to be the most successful in treating the alpha strain the patient suffers from. The side effects are moderate and the patient is highly likely to experience them, which would make drug 2 too dangerous. As a grandmother any disruption to her life caused by the side effects should not be too severe.

The patient's symptoms are not too severe and the likelihood of any drug being successful is low. Drug 3 treats both strains the patient suffers from without the risk of severe side effects. Although this risk would be low she is a busy mother, drug 2 would not offer a significantly greater success rate and therefore drug 3 is the safer and better option.

Susceptibility to side effects was high so didn't want to choose 2 because of severe side effects said is may affect quality of life - wanted the drug with least side effects/mildest because he is elderly so I feel quality of life is more important than a drug with higher success rate but severe side effects. His response likelihood is high anyway. Didn't want drug 1 because he already has nausea and even more would be horrible! Was close – nearly chose drug 2 but the severe side effects somehow put me off.

Patients current symptoms are mild so although 2 is more likely to reduce symptoms it seems more sensible to….just changed my mind –go for 2 instead of 3 because side effects are rare and patients susceptibility and change to quality of life low/unlikely. The drowsiness etc common with 3 may impair ability to look after kids etc plus 2 is for alpha and beta with patient needs treatment for.

Higher success rate than drug 3. Most useful for treating A's infection. Although side effects are moderate unlikely to interfere with A's life too much. Most useful for treating B's infection. Side effects are rare and not too harmful (except blood clotting) drug 3 would likely produce side effects that would interfere with B's daily life.

Considered the side effects of the drug and her symptoms. It is most effective drug for treating her infection. Though the side effects are moderate I took this into account, and hope that she wont be too bust while being treated even with 5 grandkids! Questions? What is her lifestyle? Infection? Level of symptoms? low side effects – good for their busy lifestyle but also this drug is effective for treating both strains of bacteria with which they are infected. Best as I can see for symptoms, lifestyle etc.

Thought to respond well to a drug but with high susceptibility to side effects, so chose one with least worst side effects and justified lower success rate because the patient is expected to respond well and the side effects were the mildest out of the 3 drugs.

Chose one with high success rate because the patient is not expected to respond as well. Severe side effects, <u>but</u> patient is not considered very susceptible to them + also they are

“very rare”. Treats both strains so more appropriate than drug 1 which only treats alpha-bacterium.

Age of patient and lifestyle (side effects) – balancing a lower likelihood of success with a lesser chance of bad side-effects (as the patient is very susceptible) – their response to drug likelihood is high, so a drug with moderate to high success rate. looking for a drug with a high success rate, as their likelihood is low. – Less concerned with side effects as not very susceptible to them (although risk of this harming QOL if they do experience them) – focus more on success rate.

The patient will have bad side effects, and therefore the drug that offers the lowest side effects therefore best. Although the chance of success is lower in drug 1, that is only small and since she is quite elderly anyway it is better that she has a good quality of life in her last few years – especially if the treatment does not work and she dies. the patient has a low likelihood to respond to the drug and therefore the treatment that offers the highest percentage survival chance is best.

Drug 1 is most useful for treating the alpha strain which is what patient A has. It has a 60% success rate. While the side effects are moderate and common and this patient is likely to be susceptible to them, they do not seem as bad as their current symptoms. (one is even the same – nausea) so I think is would be worth it if this drug works and although it’s likely to affect quality of life this will only be during treatment so this is a short term (moderate symptoms) issue to solve a long term (severe symptoms) illness.

Since patient B has the alpha AND beta strains the choice was drug 2 or 3. The side effects for drug 2 sound awful and I would not want to put someone through that! The side effects for drug 3 seem low and although the patient has a low response likelihood to treatment I think it is better and safer to try this drug first.

Appendix 7 Self versus Other Directed Rationales: A Comparison of Reasoning Styles – Study Brief

Evidence for 'Nature' ('born with aggression')	Evidence for 'Nurture' ('aggression is learned')
• Twin and Adoption studies indicate that identical twins reared separately share many inherited similarities such as aggression. • There are genetic links to various conditions that influence aggressive behaviour such as schizophrenia • There is a wealth of research demonstrating the link between hormones such as testosterone and aggression • The Biochemical theory of gender identity suggests that we acquire our gender identities through genetic and hormonal factors rather than through socialisation. E.g. males more aggressive than females. • The Psychoanalytical approach suggests that behaviour is controlled by the innate aggression and sex drives. Society restricts these drives and that these drives are controlled via the ego (moral) and the superego (selfish) • Crime statistics support gender differences in aggression – e.g. males more likely to be imprisoned for violent crime - testosterone.	• Some regard the mind as a blank slate at birth and therefore all knowledge and understanding is ascertained through life experience. • The social learning theory suggests that we learn through observation and imitation. • Behaviourists believe that human behaviour is learnt. Operant and classical conditioning techniques demonstrate that aggressive behaviour can be easily influenced by the environment. • Psychologists looking at development have explored many developmental factors, such as child rearing practices and the influence of role models – e.g. male and female role models. • Psychologists have identified the influence of other people and phenomena such as conformity and group behaviour. E.g. diminished responsibility and rioting. • Environmental Psychologists conducted a wealth of research that demonstrates the influence of the environment on people's health and social behaviour. E.g. Hot weather increases aggression.
There are also arguments for an interaction of nature and nurture. Cognitive psychology looks at innate cognitive abilities, but recognises that experience shapes these abilities.	

You will now need to develop a rationale for a decision - whether you think humans are born with an innate level of aggression or if this is learned.

A 'rationale' is an explanation of why you chose a particular side and the reasons behind your decision as well as the thought processes you went through to choose a particular side.
Using the information you have been given above as part of your answer, now decide whether you feel that people are born with an innate level of 'aggression' or whether this is learned behaviour?

Self Prompt: Please note, your rationale will not be shared.

Other Prompt: Please note, your rationale will be shared with another person to assist in their decision making.

Appendix 8 Self versus Other Directed Rationales: Knowledge Recall Test

In this section you will need to answer a few questions just to check if you do have any prior knowledge of the psychology oriented debate topic 'Nature Vs Nurture'. Please do not refer to any other sources for your answers.

1. What type of influences on human behaviour do studies of Twins demonstrate?

Inherited characteristics
Influence of parenting styles
Physical attributes
Do not know

2. What do the processes of operant and classical conditioning help to demonstrate?
That behaviour is influenced by genes
That behaviour is influenced by the environment
That behaviour is influenced by chemicals ingested
Do not know

3. What is the fundamental belief of the 'behaviourist' school of thought?

Behaviour is due to genetic influences
Behaviour is learned
Behaviour is the result of ingested chemicals
Do not know

4. What types of social mechanism may be responsible for aggressive behaviour such as rioting?

Diminished responsibility
Conformity
Groupthink phenomenon
Do not know

5. What environmental influences may be key in aggressive behaviour according to Psychologists?

Electromagnetic fields
Air pollution density
Temperature and proximity
Do not know

6. Which mental illness is most commonly referenced in research on the link between genetics and behaviour?
ADHD
Schizophrenia
Depression
Do not know

7. How does 'Social Learning Theory' suggest our behaviours arise?

From natural tendencies
Due to our physical attributes
From imitation of others
Do not know

8. What has research shown to be the most powerful influence in predicting aggression?

Genes
Ingested chemicals
Hormones
Do not know

9. What does the biochemical theory of gender identity suggest?

That gender is influenced by chemicals ingested
That gender is influenced by genes and hormones
That gender is influenced by socialisation
Do not know

10. What is the basic concept of the 'Blank slate' theory?

Most Knowledge is innate
All knowledge attained through experience
Some knowledge is innate and some from experience
Do not know

Appendix 9 Self versus Other Directed Rationales: Sample of Rationales

I believe that a person's tendency towards aggression is shaped by both nature and nurture. Genetic and hormonal factors establish a biochemical baseline which is modified, suppressed, or cultivated by social acculturation. I would also argue that the final factor is the moral agency of the individual, but this of course is more the realm of the philosopher or theologian that it is the province of the psychologist. Despite hormonal pressure, social norms, and a host of innate and external factors, individuals ultimately get to decide whether they will act out their impulse towards aggression. Ultimately, it is too reductive to say that aggression is just nature or just nurture. Why does it need to be one or the other?

I genuinely believe it is both. I think that every person is born with different traits or predispositions, whether this be aggression or other personality/behaviour facet. However there is some evidence that both childhood and adult experiences, environmental factors and access to strong role models and mediation techniques can influence what may be considered to be innate. Some people are more prone to aggressive reactions and will respond more easily to certain stimuli, some people learn behaviours from their parents or other adult carers, some learn them because they are surviving on their own with no role model or intervention. But others learn to control or change their behaviours through counselling, mediation and mentoring. We see this often in our work in primary education.

I believe that aggressive behaviour actually derives from a balance of nature and nurture. There is clearly a difference in aggression between males and females, which relates to hormonal differences, but there can be aggressive females and non-aggressive males. The truth is that there is a degree of "heritability" in any trait, as our brains are controlled by biochemistry, which is in turn controlled by our genes, however external factors such as our upbringing can also affect our biochemistry, and therefore our behaviour. It has actually been found that there is degree of heritability in every trait (although sometimes extremely small, fractions of a percent) except how many children you have and political orientation.

I would believe that both arguments hold true, as there is evidence both contradictory and free standing for both sides. Taking for example the discussion of twins: No information is given as to the additional environmental factors surrounding the separated upbringing which could well greatly influence the behaviour more than any genetic traits. On gender differences, there are many examples of people born to one gender who clearly identify with the other, to such an extent that this is a recognised condition. Ultimately I feel that a combination of factors can be taken to influence aggression: Genetic factors that serve to set a basis for a person's innate response to stimuli and learned responses that act to moderate this level up and down. Basic in-built social behaviour can then further modify the response in some situations.

Both. I think that hormones can cause mood changes which can lead to an aggressive nature. Everyone has a different level of hormones produced which may affect how an individual feels. Each individual has their own thoughts and different ways to deal with a single solution, some may be more aggressive than others. My siblings and i grew up in a very similar environment but i know that our personalities are very different. Life experiences can enhance or change the way people think, their habits and their morals in life. I agree with conformity and social behaviour from experimental studies.

I think that aggression is somewhat innate- you are born with a balance of chemicals (such as dopamine) in the brain and people with a certain make up of chemicals will be more likely to turn aggressive. Eg the likelihood of developing schizophrenia is increased if a parent has schizophrenia. However I also think that aggression can be influenced or learnt-

by conformity or copying others and that people who are born with a different balance of chemicals will still need a certain amount of input from the environment e.g. getting into the wrong crowd at school, to then realise the aggression.

As with a lot of things, the true answer is probably somewhere in the middle. I've worked with kids closely during volunteer work, they came from council supported families and we did see a high aggression rate. We had numerous kids with ADHD and whether the aggression from everyone was how they acted at home or just how they were seemed to vary. Some parents were shocked at how their kids acted and some seemed to expect it. Overall most parents seemed friendly and not aggressive (easily could be acting though). It was just odd to see kids change just because they were around aggressive kids, I guess that's an argument for nurture. Overall, I don't know, somewhere in the grey area"

"I believe that people are built with an innate level of aggression, but environmental factors determine whether they act on it. In the above table, the importance of hormones and genetics on aggression are described. These are values that can be readily and scientifically measured and hence would perhaps be more reliable indicators of aggression than environmental influences. However people are more than just chemicals and genetics, and the arguments for nurture are also convincing. Combining the information in the table with real-life experience (and aggressive parent doesn't necessarily lead to an aggressive child) led to my conclusion."

I think it is a mixture of both arguments, i.e. an interaction of nature and nurture is responsible for people's behaviour. I think the social learning theory provides strong evidence for the 'nurture argument' especially when it comes to behavioural differences between genders. Regarding the nature argument, however, I think one cannot ignore the fact that hormones do have a strong impact on our behaviour. Every healthy women, who has experienced the wrong setting of hormonal contraception gives proof to it. Therefore, I think there is truth in both arguments, although I would emphasise the nurture argument by agreeing that experiences shape innate abilities.

I think people are born with an innate level of aggression. The evidence for 'Nature' features more directly related research and less theory than does 'Nurture'. Also, it could be argued that the 'Nurture' column simply reflects the fact that a person with their innate level of aggression will naturally express this more or less depending on their current circumstances, but if any two people were in the same situation the level of expression of aggressive behaviour would always be proportional to their innate level of aggression.

I sway towards the behaviourist theories. As a primary school teacher I can see that students adapt their level of aggression to different situations. This demonstrates quite a large level of control which would perhaps not be apparent if aggression was inherited. Also aggression is often seen during group situations such as riots or due to obedience such as during war. Perhaps men are more likely to follow group norms due to the level of testosterone, however most men demonstrate a level of control when acting individually.

Appendix 10 Self versus Other Directed Rationales: Classical RST Relation Examples

For the purposes of ease of presentation the examples of relations will be grouped into the subject matter and presentational categories.

Presentational Relations

Antithesis "I see them as just one of the natural ways to express ones emotions rather than look at them as all negative side of human being (sic)" and "I think it is a mixture, not purely one or the other"

Concession "one can determine that although nurture has an influence on when aggression will be shown it does not control the level of aggression"

Evidence "I believe that aggression is an inherent part of people's nature (…) a display of aggression is programmed into our genetics"

Justify "I don't know enough about the subject matter, to be able to make an informed decision; as I would have to blindly assume the validity of each the points"

Subject Matter

Condition "If one considers the above evidence, one can determine that although nurture has an influence on when aggression will be shown…."

Evaluation "This evidence is more conclusive than a blank slate theory – which is only a theory"

Interpretation "the idea of a blank slate at birth may be slightly different for males and females with males being more likely to become more aggressive"

Non volitional cause "if not encouraged/moulded by society then it may never be present" And "people being more successful, living longer and therefore providing more children who are in turn aggressive"

Non volitional result "social learning theory and operant and classical conditioning have the biggest influence on our aggression due to parents raising the next generation to be like themselves"

Solutionhood "But in order to eliminate people with unfavourable predispositions for the happiness and wellbeing of others we need to act as if this is not true"

And "Ultimately, it is too reductive to say that aggression is just nature or just nurture. Why does it need to be one or the other?"

Mean manner "aggression can be influenced at an innate level by virtue of genetics and sex"
And "although I would emphasise the nurture argument by agreeing that experiences shape innate abilities."

Unless "such aggressive tendencies may never be exacerbated unless that individual is in an environment that channels or activates such aggression"

Elaboration "every person is born with different traits or predispositions, whether this be aggression or other personality/behaviour facet"

Multinuclear Relations
Conjunction "aggressive people can be taught how to control their aggression and likewise non aggressive people can become so."

Disjunction "I can't learn to change my eye colour as it is purely natural; nor do I naturally speak a language as it is something I have to learn"

Joint "ultimately I feel that a combination of factors can be taken to influence aggression; genetic factors that serve to set a basis for a person's innate response"

Contrast "some people are more prone to aggressive reactions (…) some learn them become they are surviving on their own"

List "the main causes are due to upbringing, aggressive role models, not learning how to control emotions and being in bad company"

Appendix 11 Self versus Other Directed Rationales: PDTB Parser Labelling Examples

Conjunction "just what I have read and what I have seen in the world"
and "I think it's a combination; some people are born with more aggressive tendencies, and certain external environmental factors can exacerbate that aggression in those people."

Alternative relation denoted by 'unless' "however such aggressive tendencies may never be exacerbated unless an individual is in an environment that channels or activates such aggression"

Synchrony is often denoted by the use of 'when' e.g. "people take the conditions or circumstances that they are surrounded by and respond aggressively when they feel ripped off"

Concession "nature clearly has an effect through hormones due to gender and genetics however this doesn't explain the vast differences in the levels of aggression between members of the same gender"

Cause "Babies are not innately aggressive and aggression tends to be found in social groups, suggesting the environmental influence is stronger" and "this is because from a young age genetic differences can determine aggression"

Condition "if a child is born in an unloving/uncaring environment I feel s/he will be more likely to commit crimes.

Contrast "I think a person's behaviour is a factor of both to a certain degree but I am leaning more towards learned" & "I believe people are born with an innate level of aggression however there are also many situations where the behaviour is learned"

Contrast "whether the aggression from everyone was how they acted at home or just how they are seemed to vary"

Instantiation "individual and subtle differences between peers of the same group can be attributed to learned behaviour for example, the differences in aggression levels between one female and another"

Restatement "overall I don't know somewhere in the grey area"
Asynchronous "but someone (…) who plays street fighter might unleash a more specialized move like a palm strike"

Appendix 12 Self versus Other Directed Rationales: HILDA Parser Labelling Examples

Attribution in the HILDA parser seems to be largely denoted with "I" "they" or "you" "suggests" "the study of epigenetics shows that the environment in which a person grows up (i.e. nurture) plays an equally important role"

Elaboration "babies are not innately aggressive and aggression tends to be found in social groups"

Joint "you are born with a balance of chemicals (such as dopamine) in the brain and people with a certain make up of chemicals will be more likely to turn aggressive"

Condition "if someone was brought up in a very calm environment with caring parents family then whatever aggression they have innately would be severely dampened"

Background "for example when looking at the riots that happened last year, I feel it more likely that these arose due to the combination of group pressure and shared anger"

Enablement "which one I believe to be the main source it would be that it is through nurture - that aggression is learned."

Explanation "hence I chose this side because I believe the evidence is more easily proven"
Contrast "hormone production is entirely in the genetic side of the argument but it is also influenced by the environment"

Manner-means "I believe that these are triggered by environmental causes, not by a predetermined genetic reason"

Cause "developmental factors can alter the way in which genes are expressed"

Same unit – denotes an element of text that belongs to the previous identified relation but may be separated by an embedded relation or text span.

Appendix 13 Self Versus Other Directed Groups: Summary Tables

		Group	Mean	*SD*	Median
State Relations	Background	SD	0.72	0.97	0
		OD	0.02	0.14	0
	Elaboration	SD	0.72	0.97	0
		OD	0.37	0.7	0
	Conjunction	SD	0.36	0.88	0
		OD	0.2	0.61	0
	Summary	SD	0	0	0
		OD	0.06	0.24	0
	Restatement	SD	0.04	0.2	0
		OD	0.02	0.14	0
	Justify	SD	0.22	0.42	0
		OD	0.33	0.47	0
Analyse Relations	Interpretation	SD	0.68	0.87	0
		OD	1	1.08	1
	Evaluation	SD	0.26	0.53	0
		OD	0.27	0.57	0
Argue Relations	Concession	SD	0.82	0.75	1
		OD	1.29	1.02	1
	Evidence	SD	1.5	1.11	1
		OD	1.55	1.16	1
	Antithesis	SD	0.12	0.48	0
		OD	0.12	0.39	0
	Contrast	SD	0.16	0.55	0
		OD	0.24	0.66	0

Table 62 SD and OD Groups: Full Classical RST Relations summary table part 1

		Group	Mean	*SD*	Median
Additional Relations	Disjunction	SD	0.04	0.28	0
		OD	0.12	0.48	0
	Condition	SD	0.08	0.27	0
		OD	0.16	0.37	0
	Solutionhood	SD	0.04	0.2	0
		OD	0.06	0.24	0
	Means	SD	0.08	0.27	0
		OD	0.12	0.33	0
	Unless	SD	0.02	0.14	0
		OD	0.02	0.14	0
	Cause	SD	0.12	0.33	0
		OD	0.12	0.33	0
	Joint	SD	0	0	0
		OD	0.12	0.48	0
	List	SD	0.08	0.4	0
		OD	0.2	0.71	0
	Result	SD	0.04	0.2	0
		OD	0.1	0.31	0

Table 63 SD and OD Groups: Full Classical RST Relations summary table part 2

		Group	Mean	*SD*	Median
'State' Relations	Conjunction	SD	0.8	0.95	1
		OD	0.94	1.14	1
	Restatement	SD	0.14	0.5	0
		OD	0.16	0.37	0
	Instantiation	SD	0.02	0.14	0
		OD	0.08	0.28	0
'Analyse' Relations	Asynchronous	SD	0.06	0.24	0
		OD	0.2	0.5	0
	Synchronous	SD	1.14	1.37	0
		OD	0.24	0.48	0
	Cause	SD	0.22	0.51	1
		OD	1	1.09	1
'Argue' Relations	Concession	SD	0.08	0.27	0
		OD	0.18	0.49	0
	Alternative	SD	0.04	0.2	0
		OD	0.1	0.31	0
	Contrast	SD	0.7	0.79	1
		OD	1.12	0.95	1
Additional Relations	Condition	SD	0.18	0.44	0
		OD	0.22	0.42	0
	Ent Relation	SD	0.46	0.71	0
		OD	0.37	0.53	0

Table 64 SD and OD Groups: Full PDTB parser relations summary table

		Group	Mean	*SD*	Median
State Relations	Elaboration	SD	4.42	3.39	3
		OD	5.35	4.02	5
	Background	SD	0.32	0.55	0
		OD	0.45	0.65	0
	Attribution	SD	1.18	1.08	1
		OD	1.73	1.48	2
	Summary	SD	0	0	0
		OD	0	0	0
Analyse Relations	Cause	SD	0.02	0.14	0
		OD	0.02	0.14	0
Argue Relations	Comparison	SD	0.02	0.14	0
		OD	0.16	0.43	0
	Contrast	SD	0.36	0.56	0
		OD	0.53	0.58	0
	Explanation	SD	0.04	0.2	0
		OD	0.08	0.28	0
Additional Relations	Joint	SD	0.64	1.01	0
		OD	0.61	0.95	0
	Condition	SD	0.18	0.44	0
		OD	0.1	0.31	0
	Enablement	SD	0.08	0.27	0
		OD	0.06	0.24	0
	Same Unit	SD	0.24	0.48	0
		OD	0.51	0.89	0

Table 65 SD and OD Groups: Full HILDA Parser relations summary table

Appendix 14 Expert Versus Novice: Summary Tables

		Group	Mean	*SD*	Median
State Relations	Elaboration	Expert	1.5	1.2	1.5
		Novice	3.67	2.2	3.5
	Conjunction	Expert	0.72	0.75	1
		Novice	0.22	0.65	0
	Summary	Expert	0.22	0.43	0
		Novice	0.28	0.67	0
	Restatement	Expert	0.06	0.24	0
		Novice	0.28	0.46	0
	Justify	Expert	0.78	0.73	1
		Novice	0.28	0.46	0
Analyse Relations	Interpretation	Expert	0.38	0.69	1
		Novice	0.72	0.96	0.5
	Evaluation	Expert	0.22	0.43	0
		Novice	1.17	1.25	1
Argue Relations	Concession	Expert	1.39	0.7	1.5
		Novice	1.56	1.2	1
	Evidence	Expert	1.89	1.5	2
		Novice	2.61	1.2	3
	Antithesis	Expert	0.44	0.62	0
		Novice	0.44	0.86	0
	Contrast	Expert	1.11	0.96	1
		Novice	0.06	0.24	0
Additional Relations	Condition	Expert	0.56	0.51	1
		Novice	0.39	0.78	0
	Means	Expert	0.06	0.24	0
		Novice	0.11	0.32	0
	Cause	Expert	0.83	0.71	1
		Novice	0.44	0.78	0
	Result	Expert	0.39	0.78	0
		Novice	0.17	0.38	0

Table 66 Full Expert and Novice groups Classical RST relations summary table.

		Group	Mean	*SD*	Median
State Relations	Elaboration	Expert	9.61	4.95	9
		Novice	10.28	3.68	9.5
	Background	Expert	0.83	0.99	0.5
		Novice	0.78	0.73	1
	Attribution	Expert	3.33	1.6	3.5
		Novice	1.94	1.26	2
	Summary	Expert	0	0	0
		Novice	0	0	0
Analyse Relations	Cause	Expert	0.11	0.32	0
		Novice	0.06	0.24	0
Argue Relations	Comparison	Expert	0.06	0.24	0
		Novice	0.06	0.24	0
	Contrast	Expert	1.17	0.99	1
		Novice	0.5	0.62	0
	Explanation	Expert	0.39	0.78	0
		Novice	0.11	0.32	0
Additional Relations	Joint	Expert	1.61	1.8	1
		Novice	1.39	1.24	1
	Condition	Expert	0.17	0.38	0
		Novice	0.33	0.49	0
	Same Unit	Expert	0.72	0.89	0.5
		Novice	1	1.19	1

Table 67 Full Expert and Novice groups HILDA relation summary table.

		Group	Mean	*SD*	Median
State Relations	Conjunction	Expert	2.55	3.15	2
		Novice	1.44	1.15	1
	Restatement	Expert	0.44	0.71	0
		Novice	0.5	0.79	0
	Instantiation	Expert	0.28	0.46	0
		Novice	0.17	0.38	0
Analyse Relations	Asynchronous	Expert	0.06	0.24	0
		Novice	0.33	0.59	0
	Synchronous	Expert	0.67	0.91	0
		Novice	0.56	0.78	0
	Cause	Expert	2.28	1.24	2
		Novice	1.94	1.8	1.5
Argue Relations	Concession	Expert	0	0	0
		Novice	0.22	0.15	0
	Alternative	Expert	0.06	0.24	0
		Novice	0.39	0.5	0
	Contrast	Expert	2.06	1.11	2
		Novice	1.33	1.28	1
Additional Relations	Condition	Expert	0.39	0.61	0
		Novice	0.5	0.51	0.5
	Ent Relation	Expert	0.11	0.32	0
		Novice	0.5	0.86	0

Table 68 Full Expert and Novice groups PDTB relations summary table.

Appendix 15 Evaluation Study: Rationale Evaluation Set

Rationale One “I feel that (most) people are born with an innate level of 'aggression' instead of being entirely learned behaviour. This is because from a young age genetic differences can determine aggression, as shown in the Twin Study. As a person ages their hormones such as Testosterone will increase, and the biochemical theory suggests gender is influenced by hormones rather than effects of the environment. Genetic conditions such as Schizophrenia also affect behaviour. Hence I chose this side because I believe the evidence is more easily proven by physical differences in a brain, rather than observed behaviour which is harder to determine and get accurate results from.”

Rationale Two "I think it is a mixture of both arguments, i.e. an interaction of nature and nurture is responsible for people's behaviour. I think the social learning theory provides strong evidence for the 'nurture argument' especially when it comes to behavioural differences between genders. Regarding the nature argument, however, I think one cannot ignore the fact that hormones do have a strong impact on our behaviour. Every healthy women, who has experienced the wrong setting of hormonal contraception gives proof to it. Therefore, I think there is truth in both arguments, although I would emphasise the nurture argument by agreeing that experiences shape innate abilities."

Rationale Three "I would believe that both arguments hold true, as there is evidence both contradictory and free standing for both sides. Taking for example the discussion of twins: No information is given as to the additional environmental factors surrounding the separated upbringing which could well greatly influence the behaviour more than any genetic traits. On gender differences, there are many examples of people born to one gender who clearly identify with the other, to such an extent that this is a recognised condition. Ultimately I feel that a combination of factors can be taken to influence aggression: Genetic factors that serve to set a basis for a persons innate response to stimuli and learned responses that act to moderate this level up and down. Basic in-built social behaviour can then further modify the response in some situations."

Rationale Four "More testosterone, more likely to get angry. However can be controlled by learnt behaviour. Nature provides the bomb, nurture sets it off. Environment can cause people to become more likely to unleash anger/aggression by "shortening the fuse" however still requires a spark to ignite it. I believe some people are born with a 'larger explosion' as it were, i.e. when they get angry they get really angry and aggressive, but the cause of the anger and the outlet of the aggression are nurture, i.e. you can choose what to hit but not necessarily how hard you hit it."

Appendix 16 Quality Level Descriptive Data

	Quality Levels									
	1 (N=12)		2 (N=37)		3 (N=35)		4 (N=22)		5 (N=9)	
Toulmin Elements	**M**	***SD***	**M**	***SD***	**Mean**	***SD*.**	**M**	***SD***	**M**	***SD***
Claims	1.33	0.49	1.62	0.79	1.91	0.92	2.22	1.15	2.64	1.02
Backing	1.00	0.60	2.27	1.47	3.54	2.41	4.95	2.48	5.72	4.41
Counter Claims	0.08	0.29	0.24	0.49	0.49	0.51	0.91	0.61	0.91	0.30
Rebuttals	0.08	0.29	0.08	0.27	0.54	0.61	0.91	0.53	1.55	0.82

Table 69 Means for Toulmin Elements within the rationales categorised at each quality level.

	Quality Levels									
	1 (N=12)		2 (N=37)		3 (N=37)		4 (N=22)		5 (N=9)	
Classical RST	**M**	***SD***	**M**	***SD***	**M**	***SD***	**M**	***SD***	**M**	***SD***
Conjunction	0.00	0.00	0.38	0.79	0.23	0.81	0.55	0.80	0.64	0.81
Restatement	0.00	0.00	0.00	0.00	0.08	0.28	0.05	0.21	0.00	0.00
Elaboration	0.08	0.29	0.51	0.77	0.65	0.86	1.09	1.23	1.55	1.37
Summary	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.39	0.27	0.47
Justify	0.17	0.39	0.22	0.42	0.46	0.66	0.45	0.51	0.45	0.52
Background	0.00	0.00	0.41	0.72	0.31	0.63	0.50	1.10	0.00	0.00
Interpretation	0.25	0.45	0.38	0.64	1.03	0.92	1.09	1.02	1.63	1.21
Evaluation	0.08	0.29	0.27	0.51	0.22	0.55	0.36	0.66	0.27	0.47
Evidence	0.50	0.67	1.30	1.02	1.69	0.99	2.14	1.17	2.27	1.74
Antithesis	0.00	0.00	0.16	0.55	0.09	0.37	0.27	0.55	0.45	0.52
Concession	0.42	0.51	0.68	0.67	1.26	0.66	1.68	1.17	1.73	0.65
Contrast	0.00	0.00	0.16	0.55	0.31	0.71	0.45	0.86	1.18	0.98
Cause	0.00	0.00	0.05	0.23	0.20	0.41	0.45	0.60	0.73	0.79
Condition	0.00	0.00	0.14	0.35	0.06	0.24	0.32	0.48	0.73	0.47
Result	0.00	0.00	0.11	0.31	0.03	0.17	0.36	0.73	0.09	0.30
Disjunction	0.00	0.00	0.00	0.00	0.06	0.34	0.09	0.43	0.36	0.81

Table 70 Means for Classical RST relations within the rationales categorised at each quality level.

	Quality Levels									
	1 (N=12)		2 (N=37)		3 (N=37)		4 (N=22)		5 (N=9)	
HILDA	**M**	***SD***	**M**	***SD***	**M**	***SD***	**M**	***SD***	**M**	***SD***
Contrast	0.08	0.29	0.49	0.56	0.51	0.61	0.91	1.02	0.73	0.65
Condition	0.00	0.00	0.19	0.40	0.08	0.36	0.14	0.35	0.36	0.51
Explanation	0.00	0.00	0.05	0.23	0.11	0.32	0.09	0.29	0.45	0.93
Enablement	0.00	0.00	0.08	0.28	0.11	0.32	0.00	0.00	0.00	0.00
Comparison	0.00	0.00	0.11	0.39	0.09	0.28	0.09	0.29	0.09	0.30
Background	0.00	0.00	0.24	0.43	0.49	0.70	0.73	0.83	1.00	0.89
Cause	0.00	0.00	0.00	0.00	0.03	0.17	0.09	0.29	0.09	0.30
Elaboration	1.83	1.64	3.41	2.22	4.89	2.62	9.45	3.58	11.72	5.50
Attribution	0.42	0.67	0.97	0.90	2.03	1.58	3.05	1.40	2.27	1.49
Joint	0.25	0.45	0.46	0.93	0.91	1.38	0.91	0.92	1.73	1.79
Same unit	0.00	0.00	0.08	0.28	0.46	0.66	0.77	0.81	1.27	1.35

Table 71 Means for HILDA Parser relations within the rationales categorised at each quality level.

	Quality Levels									
	1 (N=12)		2 (N=37)		3 (N=37)		4 (N=22)		5 (N=9)	
PDTB	**M**	***SD***	**M**	***SD***	**M**	***SD***	**M**	***SD***	**M**	***SD***
Contrast	0.25	0.45	0.57	0.60	1.29	0.75	1.77	1.41	1.67	0.87
Conjunction	0.33	0.65	0.73	0.99	1.00	0.97	1.55	1.37	3.11	2.98
Restatement	0.00	0.00	0.11	0.31	0.17	0.57	0.36	0.58	0.44	0.53
Cause	0.33	0.65	0.86	1.16	1.23	1.03	2.05	1.43	2.33	1.66
Instantiation	0.00	0.00	0.00	0.00	0.06	0.24	0.23	0.43	0.22	0.44
Alternative	0.00	0.00	0.00	0.00	0.06	0.24	0.14	0.35	0.33	0.50
Asynchronous	0.00	0.00	0.05	0.23	0.09	0.28	0.36	0.66	0.11	0.33
Synchronous	0.00	0.00	0.24	0.43	0.11	0.32	0.68	0.89	0.67	0.87
Concession	0.00	0.00	0.11	0.31	0.09	0.28	0.18	0.50	0.22	0.67
Condition	0.00	0.00	0.19	0.40	0.14	0.43	0.45	0.60	0.56	0.53
Ent Rel	0.17	0.39	0.41	0.64	0.37	0.49	0.36	0.73	0.44	0.73

Table 72 Means for PDTB Parser relations within the rationales categorised at each quality level.

13 References

Azar, M. (1999). Argumentative text as rhetorical structure: An application of rhetorical structure theory. *Argumentation, 13*(1), 97-114.

Baghaei, N., Mitrovic, A., & Irwin, W. (2007). Supporting collaborative learning and problem-solving in a constraint-based CSCL environment for UML class diagrams. *International Journal of Computer-Supported Collaborative Learning, 2*(2-3), 159-190.

Belgiorno, F., De Chiara, R., Manno, I., Overdijk, M., Scarano, V., & van Diggelen, W. (2008). Face to face cooperation with CoFFEE *Times of Convergence. Technologies Across Learning Contexts* (pp. 49-57): Springer.

Bender, R., & Lange, S. (1999). Multiple test procedures other than Bonferroni's deserve wider use. *BMJ: British Medical Journal, 318*(7183), 600.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*: Lawrence Erlbaum Associates Hillsdale, NJ.

Berthold, K., Eysink, T. H., & Renkl, A. (2009). Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science, 37*(4), 345-363.

Bex, F., Budzynska, K., & Walton, D. (2012). Argumentation and explanation in the context of dialogue. *Explanation-aware Computing ExaCt 2012, 9*, 6.

Bex, F., Lawrence, J., Snaith, M., & Reed, C. (2013). Implementing the argument web. *Communications of the ACM, 56*(10), 66-73.

Bex, F., Van den Braak, S., Van Oostendorp, H., Prakken, H., Verheij, B., & Vreeswijk, G. (2007). Sense-making software for crime investigation: how to combine stories and arguments? *Law, Probability and Risk, 6*(1-4), 145-168.

Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction, 13*(2), 221-252.

Buckingham Shum, S., & Okada, A. (2008). Knowledge cartography for open sensemaking communities. *Journal of Interactive Media in Education, 2008*(1).

Burge, J. E., & Brown, D. C. (2003). *Rationale support for maintenance of large scale systems.* Paper presented at the ELISA workshop.

Burge, J. E., & Brown, D. C. (2006). Rationale-based support for software maintenance *Rationale management in software engineering* (pp. 273-296): Springer.

Carlson, L., Marcu, D., & Okurowski, M. E. (2003). *Building a discourse-tagged corpus in the framework of rhetorical structure theory*: Springer.

Carr, C. S. (2003). Using computer supported argument visualization to teach legal argumentation *Visualizing argumentation* (pp. 75-96): Springer.

Chamberland, M., St-Onge, C., Setrakian, J., Lanthier, L., Bergeron, L., Bourget, A., . . . Rikers, R. (2011). The influence of medical students' self-explanations on diagnostic performance. *Medical education, 45*(7), 688-695.

Chi, M. T. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in instructional psychology, 5*, 161-238.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989a). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(2), 145-182.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989b). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(2), 145-182.

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477.

Chittleborough, P., & Newman, M. E. (1993). Defining the Term" Argument". *Informal Logic, 15*(3).

Coleman, E. B. (1998). Using explanatory knowledge during collaborative problem solving in science. *Journal of the Learning Sciences, 7*(3-4), 387-427.

Conklin, J., & Begeman, M. L. (1988). gIBIS: A hypertext tool for exploratory policy discussion. *ACM Transactions on Information Systems (TOIS), 6*(4), 303-331.

Cooper, M. M., Cox Jr, C. T., Nammouz, M., Case, E., & Stevens, R. (2008). An assessment of the effect of collaborative groups on students' problem-solving strategies and abilities. *Journal of Chemical Education, 85*(6), 866.

Corbel, A., Jaillon, P., Serpaggi, X., Baker, M., Quignard, M., Lund, K., & Séjourné, A. (2003). DREW: Un outil Internet pour créer des situations d'apprentissage coopérant. *Desmoulins, Marquet, & Bouhineau (Eds.), Eiah2003 environnements informatiques pour l'apprentissage humain*, 109-113.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671-684.

Crammond, J. G. (1998). The uses and complexity of argument structures in expert and student persuasive writing. *Written Communication, 15*(2), 230-268.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics, 69*(9), 970-977.

Crowhurst, M. (1996). Teaching and learning argumentative writing in the middle school years. *Perspectives on written argument*, 57-72.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). *A framework and graphical development environment for robust NLP tools and applications.* Paper presented at the ACL.

Dewey, J. (1910). *How We Think*. Boston: Heath and Co.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of research in education, 32*(1), 268-291.

Easterday, M. W., Aleven, V., & Scheines, R. (2007). 'Tis Better to Construct than to Receive? The Effects of Diagram Tools on Causal Reasoning. *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 158*, 93.

Feather, N. (1968). Change in Confidence Following Success or Failure as a Predictor of Subsequent Performance. *Journal of Personality and Social Psychology, 9*(1), 38.

Felton, M., & Herko, S. (2004). From Dialogue to Two-Sided Argument: Scaffolding Adolescents' Persuasive Writing. *Journal of Adolescent & Adult Literacy, 47*(8), 672-683. doi: 10.2307/40016901

Felton, M., & Kuhn, D. (2001). The development of argumentive discourse skill. *Discourse Processes, 32*(2-3), 135-153.

Feng, V. W., & Hirst, G. (2012). *Text-level discourse parsing with rich linguistic features*. Paper presented at the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, Jeju Island, Korea.

Ferretti, R. P., MacArthur, C. A., & Dowdy, N. S. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology, 92*(4), 694.

Goodman, B. A., Linton, F. N., Gaimari, R. D., Hitzeman, J. M., Ross, H. J., & Zarrella, G. (2005). Using dialogue features to predict trouble during collaborative learning. *User Modeling and User-Adapted Interaction, 15*(1-2), 85-134.

Gordon, T. F., & Karacapilidis, N. (1997). *The Zeno argumentation framework.* Paper presented at the Proceedings of the 6th international conference on Artificial intelligence and law.

Gordon, T. F., Prakken, H., & Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence, 171*(10), 875-896.

Green, N. (2007). A study of argumentation in a causal probabilistic humanistic domain: Genetic counseling. *International Journal of Intelligent Systems, 22*(1), 71-93. doi: 10.1002/int.20190

Green, N. (2013). Towards Automated Analysis of Student Arguments. In H. C. Lane, K. Yacef, J. Mostow & P. Pavlik (Eds.), *Artificial Intelligence in Education* (Vol. 7926, pp. 591-594): Springer Berlin Heidelberg.

Green, N. L. (2010). Representation of argumentation in text with rhetorical structure theory. *Argumentation, 24*(2), 181-196.

Guenther, C. L., & Alicke, M. D. (2008). Self-enhancement and belief perseverance. *Journal of Experimental Social Psychology, 44*(3), 706-712.

Hair, D. C. (1991). Legalese: A legal argumentation tool. *ACM SIGCHI Bulletin, 23*(1), 71-74.

Hausmann, R., Chi, M. T., & Roy, M. (2004). *Learning from collaborative problem solving: An analysis of three hypothesized mechanisms.* Paper presented at the 26nd annual conference of the Cognitive Science society.

Hausmann, R., & VanLehn, K. (2007). *Self-explaining in the classroom: Learning curve evidence.* Paper presented at the Proceedings of the 29th Annual Cognitive Science Society.

Hausmann, R. G., van de Sande, B., & VanLehn, K. (2008). Are self-explaining and coached problem solving more effective when done by pairs of students than alone? *arXiv preprint arXiv:0805.4223*.

Heath, C., & Gonzalez, R. (1995). Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making. *Organizational Behavior and Human Decision Processes, 61*(3), 305-326.

Hernault, H., Prendinger, H., & Ishizuka, M. (2010). HILDA: a discourse parser using support vector machine classification. *Dialogue & Discourse, 1*(3).

Hirst, G., & hernault, H. (2015). HILDA Text Parser Retrieved 13/06/2013, from http://t2d.globallabproject.net/

Hoffman, K., & Elwin, C. (2004). The relationship between critical thinking and confidence in decision making.

Hollingshead, A. B., & McGrath, J. E. (1995). Computer-assisted groups: A critical review of the empirical research. *Team effectiveness and decision making in organizations*, 46-78.

Israel, J., & Aiken, R. (2007). Supporting collaborative learning with an intelligent web-based system. *International Journal of Artificial Intelligence in Education, 17*(1), 3-40.

Jacoby, L. L. (1983). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning and Verbal Behavior, 22*(5), 485-508.

Jermann, P., & Dillenbourg, P. (2003). Elaborating new arguments through a CSCL script *Arguing to learn* (pp. 205-226): Springer.

Johnson, B. T., & Eagly, A. H. (1989). Effects of involvement on persuasion: A meta-analysis. *Psychological bulletin, 106*(2), 290.

Joiner, R., Jones, S., & Doherty, J. (2008). Two studies examining argumentation in asynchronous computer mediated communication. *International Journal of Research & Method in Education, 31*(3), 243-255. doi: 10.1080/17437270802416848

Karacapilidis, N., & Papadias, D. (2001). Computer supported argumentation and collaborative decision making: the HERMES system. *Information systems, 26*(4), 259-277.

Karacapilidis, N., & Tzagarakis, M. (2009). Supporting Argumentative Collaboration in Communities of Practice: The CoPe_it! Approach. *Solutions and Innovations in Web-Based Technologies for Augmented Learning: Improved Platforms, Tools and Applications, IGI Global, Hershey, PA, USA*, 245-257.

Kim, J., Shaw, E., Ravi, S., Tavano, E., Arromratana, A., & Sarda, P. (2008). *Scaffolding on-line discussions with past discussions: An analysis and pilot study of PedaBot.* Paper presented at the Intelligent Tutoring Systems.

Klein, M., & Iandoli, L. (2008). Supporting Collaborative Deliberation Using a Large-Scale Argumentation System: The MIT Collaboratorium.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological bulletin, 110*(3), 499.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107.

Kuhn, D. (1991). *The skills of argument*: Cambridge University Press.

Kuhn, D., Shaw, V., & Felton, M. (1997). Effects of dyadic interaction on argumentive reasoning. *Cognition and Instruction, 15*(3), 287-315.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child development, 74*(5), 1245-1260.

Kuhn, L., & Reiser, B. (2005). *Students constructing and defending evidence-based scientific explanations.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Dallas, TX.

Kumar, R., Rosé, C. P., Wang, Y.-C., Joshi, M., & Robinson, A. (2007). Tutorial dialogue as adaptive collaborative learning support. *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 158*, 383.

Langer, J. A., & Applebee, A. N. (1987). *How Writing Shapes Thinking: A Study of Teaching and Learning. NCTE Research Report No. 22*: ERIC.

Lin, H.-s., Hong, Z.-R., & Lawrenz, F. (2012a). Promoting and Scaffolding Argumentation through Reflective Asynchronous Discussions. *Computers & Education, 59*(2), 378-384.

Lin, X., & Lehman, J. D. (1999). Supporting learning of variable control in a computer-based biology environment: Effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching, 36*(7), 837-858.

Lin, X., Newby, T. J., Glenn, N. G., & Lafayette, W. (1994). To Promote Far Transfer Problem Solving. *DOCUMENT RESUME*, 462.

Lin, Z., Ng, H. T., & Kan, M.-Y. (2012b). A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 1-34.

Lin, Z., Ng, H. T., & Kan, M.-Y. (2015). PDTB-Style End to End Discourse Parser Retrieved 12/06/2013, from http://wing.comp.nus.edu.sg/~linzihen/parser/

Loll, F., Pinkwart, N., Scheuer, O., & McLaren, B. M. (2012). How tough should it be? Simplifying the development of argumentation systems using a configurable platform. *Educational Technologies for Teaching Argumentation Skills*, 169-197.

Loui, R. P., Norman, J., Altepeter, J., Pinkard, D., Craven, D., Linsday, J., & Foltz, M. (1997). *Progress on Room 5: A testbed for public interactive semi-formal legal argumentation.* Paper presented at the Proceedings of the 6th international conference on Artificial intelligence and law.

Lowrance, J. D., Garvey, T. D., & Strat, T. M. (2008). A framework for evidential-reasoning systems *Classic Works of the Dempster-Shafer Theory of Belief Functions* (pp. 419-434): Springer.

Lu, J., Chiu, M. M., & Law, N. W. (2011). Collaborative argumentation and justifications: A statistical discourse analysis of online discussions. *Computers in Human Behavior, 27*(2), 946-955.

MacLean, A., Bellotti, V., & Shum, S. (1993). Developing the design space with design space analysis. *NORTH HOLLAND STUDIES IN TELECOMMUNICATION*, 197-197.

MacLean, A., Young, R. M., Bellotti, V. M., & Moran, T. P. (1991). Questions, options, and criteria: Elements of design space analysis. *Human–computer interaction, 6*(3-4), 201-250.

Mann, B., & Taboada, M. (2015). Rhetorical Structure Theory Retrieved 10/06/2013, from http://www.sfu.ca/rst/01intro/definitions.html

Mann, B., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3), 243-281.

Mann, B., & Thompson, S. A. (1992). *Discourse description: Diverse linguistic analyses of a fund-raising text* (Vol. 16): John Benjamins Publishing.

Mann, B., & Thompson, S. A. (2002). *Two views on Rhetorical Structure Theory.* Paper presented at the Proceedings of the 10th Annual Meeting of the Society for Text and Discourse.

Marshall, C. C., Halasz, F. G., Rogers, R. A., & Janssen Jr, W. C. (1991). *Aquanet: a hypertext tool to hold your knowledge in place.* Paper presented at the Proceedings of the third annual ACM conference on Hypertext.

Marttunen, M. (1998). Electronic mail as a forum for argumentative interaction in higher education studies. *Journal of Educational Computing Research, 18*(4), 387-405.

Mcalister, S., Ravenscroft, A., & Scanlon, E. (2004). Combining interaction and context design to support collaborative argumentation using a tool for synchronous CMC. *Journal of Computer Assisted Learning, 20*(3), 194-204.

McLaren, B. M., Scheuer, O., De Laat, M., Hever, R., De Groot, R., & Rosé, C. P. (2007). Using machine learning techniques to analyze and support mediation of student e-discussions. *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 158*, 331.

Mentis, H. M., Bach, P. M., Hoffman, B., Rosson, M. B., & Carroll, J. M. (2009). *Development of decision rationale in complex group decision making.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Miller, P. M., & Fagley, N. S. (1991). The effects of framing, problem variations, and providing rationale on choice. *Personality and Social Psychology Bulletin, 17*(5), 517-522.

Munneke, L., van Amelsvoort, M., & Andriessen, J. (2003). The role of diagrams in collaborative argumentation-based learning. *International Journal of Educational Research, 39*(1), 113-131.

Nussbaum, E. M. (2011). Argumentation, dialogue theory, and probability modeling: Alternative frameworks for argumentation research in education. *Educational Psychologist, 46*(2), 84-106.

Nussbaum, E. M., Winsor, D., Aqui, Y., & Poliquin, A. (2007). Putting the pieces together: Online argumentation vee diagrams enhance thinking during discussions. *International Journal of Computer-Supported Collaborative Learning, 2*(4), 479-500. doi: 10.1007/s11412-007-9025-1

Okumus, S., & Unal, S. (2012). The Effects of Argumentation Model on Students' Achievement and Argumentation Skills in Science. *Procedia-Social and Behavioral Sciences, 46*, 457-461.

Osborne, J., Erduran, S., & Simon, S. (2004a). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994-1020.

Osborne, J., Erduran, S., & Simon, S. (2004b). Ideas, evidence and argument in science (IDEAS) project.

Petty, R. E., & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology, 46*(1), 69.

Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. (2006). *Toward legal argument instruction with graph grammars and collaborative filtering techniques.* Paper presented at the Intelligent Tutoring Systems.

Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. (2007). Evaluating legal argument instruction with graphical representations using largo. *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 158*, 101.

Ploetzner, R., Dillenbourg, P., Preier, M., & Traum, D. (1999). Learning by explaining to oneself and to others. *Collaborative learning: Cognitive and computational approaches*, 103-121.

Ranney, M., & Schank, P. (1998). Toward an integration of the social and the scientific: Observing, modeling, and promoting the explanatory coherence of reasoning. *Connectionist models of social reasoning and social behavior*, 245-274.

Ravenscroft, A., & McAlister, S. (2008). Investigating and promoting educational argumentation: towards new digital practices. *International Journal of Research & Method in Education, 31*(3), 317-335.

Reed, C., & Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools, 13*(04), 961-979.

Reed, C., & Walton, D. (2007). Argumentation schemes in dialogue. *Dissensus and the Search for Common Ground (Proceedings of OSSA 2007)*.

Regli, W. C., Hu, X., Atwood, M., & Sun, W. (2000). A survey of design rationale systems: approaches, representation, capture and retrieval. *Engineering with computers, 16*(3-4), 209-235.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*(1), 1-29.

Robertson, J., Good, J., & Pain, H. (1998). BetterBlether: The design and evaluation of a discussion tool for education. *International Journal of Artificial Intelligence in Education, 9*(3-4), 219-236.

Rolf, B., & Magnusson, C. (2002). *Developing the art of argumentation. A software approach.* Paper presented at the Proceedings of the 5th International Conference on Argumentation. International Society for the Study of Argumentation (ISSA-2002).

Rosé, C., Bhembe, D., Siler, S., Srivastava, R., & VanLehn, K. (2003). The role of why questions in effective human tutoring. *Proceedings of the 11th International Conference on AI in Education*, 55-62.

Rottman, B. M., & Keil, F. C. (2011). What matters in scientific explanations: Effects of elaboration and content. *Cognition, 121*(3), 324-337.

Rowe, G., Macagno, F., Reed, C., & Walton, D. (2006). Araucaria as a tool for diagramming arguments in teaching and studying philosophy. *Teaching Philosophy, 29*(2), 111-124.

Roy, M., & Chi, M. T. (2005). The self-explanation principle in multimedia learning. *The Cambridge handbook of multimedia learning*, 271-286.

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*(1), 23-55.

Schank, R. C. (1986). *Explanation patterns: Understanding mechanically and creatively*: Psychology Press.

Schellens, T., Van Keer, H., De Wever, B., & Valcke, M. (2007). Scripting by assigning roles: Does it improve knowledge construction in asynchronous discussion groups? *International Journal of Computer-Supported Collaborative Learning, 2*(2-3), 225-246.

Scheuer, O., Loll, F., Pinkwart, N., & McLaren, B. M. (2010). Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning, 5*(1), 43-102.

Schneider, D. C., Voigt, C., & Betz, G. (2007). *ArguNet—a software tool for collaborative argumentation analysis and research.* Paper presented at the 7th Workshop on Computational Models of Natural Argument (CMNA VII).

Schulenberg, J. L. (2007). Analysing Police Decision-Making: Assessing the Application of a Mixed-Method/Mixed-Model Research Design. *International Journal of Social Research Methodology, 10*(2), 99-119.

Schwarz, B. B., & Glassner, A. (2007). The role of floor control and of ontology in argumentative activities with discussion-based tools. *International Journal of Computer-Supported Collaborative Learning, 2*(4), 449-478.

Schworm, S., & Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *Journal of Educational Psychology, 99*(2), 285.

Scoufis, M. U. o. W. S. N. C. T., & Writing, N. (1999). On the track critical thinking in assignment writing. Nepean, N.S.W.; [Bendigo, Vic.]: Critical Thinking & Writing Network, University of Western Sydney, Nepean ; Video Education Australasia [distributor].

Selvin, A., Buckingham Shum, S., Seirhuis, M., Conklin, J., Zimmerman, B., Palus, C., . . . Motta, E. (2001). Compendium: Making meetings into knowledge events.

Sherif, M., & Sherif, C. M. (1967). Attitudes as the individual's own categories: The social judgment involvement approach to attitude and attitude change. In C. W. Sherif & M. Sherif (Eds.), *Attitude, ego-involvement, and change.* (pp. 105–139). Chicago: Rand McNally.

Shum, S. B., & Hammond, N. (1994). Argumentation-based design rationale: what use at what cost? *International Journal of Human-Computer Studies, 40*(4), 603-652.

Sieck, W., & Yates, J. F. (1997). Exposition effects on decision making: Choice and confidence in choice. *Organizational Behavior and Human Decision Processes, 70*(3), 207-219.

Simon, S., Erduran, S., & Osborne, J. (2006). Learning to teach argumentation: Research and development in the science classroom. *International Journal of Science Education, 28*(2-3), 235-260.

Simon, S., & Johnson, S. (2008). Professional learning portfolios for argumentation in school science. *International Journal of Science Education, 30*(5), 669-688.

Soller, A. (2004). Computational modeling and analysis of knowledge sharing in collaborative distance learning. *User Modeling and User-Adapted Interaction, 14*(4), 351-381.

Soricut, R., & Marcu, D. (2003). *Sentence level discourse parsing using syntactic and lexical information.* Paper presented at the Proceedings of the 2003 Conference of the North

American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1.

Stein, N. L., & Miller, C. A. (1993). The development of memory and reasoning skill in argumentative contexts: Evaluating, explaining, and generating evidence.

Suthers, D., Connelly, J., Lesgold, A., Paolucci, M., Toth, E., Toth, J., & Weiner, A. (2001). Representational and advisory guidance for students learning scientific inquiry. *Smart machines in education: The coming revolution in educational technology*, 7-35.

Suthers, D., Weiner, A., Connelly, J., & Paolucci, M. (1995). *Belvedere: Engaging students in critical discussion of science and public policy issues.* Paper presented at the Proceedings of the 7th World Conference on Artificial Intelligence in Education.

Taboada, M. (2006). Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics, 38*(4), 567-592.

Tintarev, N., & Masthoff, J. (2007). *Effective explanations of recommendations: user-centered design.* Paper presented at the Proceedings of the 2007 ACM conference on Recommender systems.

Toulmin, S. (1958). The Uses of Argument. . *Cambridge: Cambridge UP*.

Tsovaltzi, D., Rummel, N., McLaren, B. M., Pinkwart, N., Scheuer, O., Harrer, A., & Braun, I. (2010). Extending a virtual chemistry laboratory with a collaboration script to promote conceptual learning. *International Journal of Technology Enhanced Learning, 2*(1), 91-110.

van den Braak, S. W., & Vreeswijk, G. A. (2006). AVER: Argument visualization for evidential reasoning. *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 152*, 151.

Van Eemeren, F. H., Grootendorst, R., Johnson, R. H., Plantin, C., & Willard, C. A. (2013). *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*: Routledge.

Van Gelder, T. (2002). Argument mapping with reason! able. *The American Philosophical Association Newsletter on Philosophy and Computers, 2*(1), 85-90.

Van Gelder, T. (2007). The rationale for Rationale™. *Law, Probability and Risk, 6*(1-4), 23-42.

VanLehn, K., & Jones, R. M. (1993). What mediates the self-explanation e ect? Knowledge gaps, schemas or analogies? *Ann Arbor, 1001*, 48109-42110.

Verheij, B. (2003). Artificial argument assistants for defeasible argumentation. *Artificial Intelligence, 150*(1), 291-324.

Von Aufschnaiter, C., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching, 45*(1), 101-131.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the development of children*, 34-41.

Walton, D. (2000). The place of dialogue theory in logic, computer science and communication studies. *Synthese, 123*(3), 327-346.

Weerasinghe, A., & Mitrovic, A. (2006). Facilitating deep learning through self-explanation in an open-ended domain. *International journal of Knowledge-based and Intelligent Engineering systems, 10*(1), 3-19.

Wells, S., Gourlay, C., & Reed, C. (2009). Argument blogging. *Proceedings of Computational Models of Natural Argument, CMNA*.

Williams, J., Lombrozo, T., & Rehder, B. (2010). *Why does explaining help learning? Insight from an explanation impairment effect.* Paper presented at the Proceedings of the 32nd Annual Conference of the Cognitive Science Society.

Wolfe, C. R. (2011). Some empirical qualifications to the arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(02), 92-93.

Wolfe, M. B., & Goldman, S. R. (2005). Relations between adolescents' text processing and reasoning. *Cognition and Instruction, 23*(4), 467-502.

Woolf, B. P., Murray, T., Marshall, D., Dragon, T., Kohler, K., Mattingly, M., . . . Sammons, J. (2005). *Critical Thinking Environments for Science Education.* Paper presented at the AIED.

Wyner, A., Schneider, J., Atkinson, K., & Bench-Capon, T. J. (2012). *Semi-Automated Argumentative Analysis of Online Product Reviews.* Paper presented at the COMMA.

Xiao, L. (2013a). Do members converge to similar reasoning styles in teamwork? A study of shared rationales in small team activities.

Xiao, L. (2013b). The effects of a shared free form rationale space in collaborative learning activities. *Journal of Systems and Software, 86*(7), 1727-1737.

Yeh, K.-H., & She, H.-C. (2010). On-line synchronous scientific argumentation learning: Nurturing students' argumentation ability and conceptual change in science context. *Computers & Education, 55*(2), 586-602.

Zhang, Y., Luo, X., Li, J., & Buis, J. J. (2013). A semantic representation model for design rationale of products. *Advanced Engineering Informatics, 27*(1), 13-26.